Facultad de Ciencias Sociales
Escuela de Psicología

MYTHS, TESTS AND GAMES: CULTURAL ROOTS AND CURRENT ROUTES OF ARTIFICIAL INTELLIGENCE

By

Roberto Musa Giuliano

Thesis submitted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the degree of Doctor in Psychology.

Advisor:
Carlos Cornejo Alarcón

Thesis Committee:
Francisco Ceric Garrido
Ricardo Rosas Díaz

September, 2019
Santiago, Chile

To my parents,
Tulia Giuliano Burrows and Roberto Musa Cavallé

Words of Thanks

What pleasant news to learn that the health benefits of gratitude are legion, or so wellness gurus assure us, since the struggling writer, having at last ceased in his toil, is bound by a sacred moral duty to attempt the issuance of thanks where they are due, no matter how little faith he might place in his ability to be either fair or exhaustive. (This would certainly account for the prefatory *remerciement* slowly inching its way into becoming a full-fledged literary genre in its own right.) A beneficial side-effect of trying to be comprehensive in acknowledging our incurred debts, is that all those readers not mentioned therein, but who nevertheless move through things page by page and from start to finish, shall arrive at the actual flesh of the text and find it delightfully entertaining by contrast to the dilatory list of names entirely unfamiliar.

For their role (which I hope they don't end up coming to abhor) in the course of my intellectual meanderings I must mention the following formative figures: I thank Enrique Vega for instilling in me an early love of philosophical thinking, and illustrating the subtle nuances that distinguish peripatetic discussions from vagabonding. From Franco Simonetti I learned the immense value of independence in carving up one's own areas of theoretical interest and not letting traditional disciplinary boundaries constrain our curiosity. I'm glad to have succeeded at Wason's selection task in one of his classes, as that event led to many talks on rationality and the hazards it faces and eventually to a friendship thanks to which I've grown enormously over the years. I feel grateful for the encouragement and kindness of Roberto González as well as his welcoming me into his research group, where I saw how research gains from being conducted by a friendly and committed team. To Ricardo Rosas I owe the living example that psychological research and a love for literature can be both fruitfully cultivated in unison as well as an invitation to a breakfast where Paul Auster delivered an address on writing; an invitation so wonderful and unexpected that it seemed like a veritable *Deus Ex Machina* and I shall never forget it. I thank Francisco Ceric for being a model as a masterful lecturer, making the details of brain physiology come fully alive even for someone lacking a solid biological background (and for mercifully, not spoiling the ending of too many series and movies in the process).

The School of Psychology at Pontificia Universidad Católica is an institutional home I'm happy to belong to and where I have always felt welcome. Were it not for Bernardita Villarroel's constant help

and patience above and beyond the call of duty, I hesitate to even speculate if that would still be the case. Likewise, I'm happy to have been able to be a part of the Language, Interaction and Phenomenology Laboratory (LIF) and to have had the chance of contributing to the valuable work carried out therein.

I will forever be indebted to Diego Cosmelli who invited me to help him organize the lectures that Douglas Hofstadter delivered in Chile in 2014, which eventually led to my spending a year with his research group at Indiana University as a visiting scholar. Francisco Claro's help was also critical in this regard, and I will never forget how he so deeply and positively influenced the course of my life. I'm thankful to Christian Berger for vouching for this stay abroad during the course of my doctorate, from which this thesis benefitted greatly. Douglas Hofstadter had long been one of my intellectual heroes and the opportunity to work closely with him not only immensely enriched my outlook on Artificial Intelligence and (more importantly) human thinking, but made me realize that his generosity and warmth outshine even his astonishing stature as a writer and thinker.

My friend José Ignacio Contreras once wrote that friendship is the most inexhaustible of libraries and I have had ample chance to verify the truth of his statement over these past years from having benefited in refining my thinking on machine intelligence from deep conversations with Kael Becerra, Marcelo Bova, Diego Carrasco, David Carré, Felipe Cuadra, Ana María Cvitanic, Pedro González, Conrado Hayler, Andrea Lobos, Sofía Ormazabal and Juan José Valenzuela. Special thanks go to Esteban Hurtado, who orchestrated my first meeting, long ago, with the theoretical implications of Turing machines and also taught me what small amount of coding I'm capable of today (an amount which I never lose the hope of increasing).

To my advisor, Carlos Cornejo, I owe more than I can articulately convey. He has been a fundamental inspiration as a scholar and as a human being, and I am glad to have had the great fortune of learning from him, first as a teaching assistant, then as a research assistant and now as his doctoral student, ever since the day that as an undergraduate I chanced to ask him what the difference was between the words 'Seele' and 'Geist', a question that having underwent several metamorphoses, to this day I still pursue. He is the best embodiment I know of that C.P. Snow's divide between the two cultures can be bridged and his research offers constant proof that if psychology is not aware of its past it will be blind to its future.

My parents have shown me love, faith and support beyond belief, and if there is one thing to fault them with it is only not having provided me with one of those unhappy childhoods that crowd wisdom credits as an indispensable requirement for talent in all things literary. My brothers, Pablo and Renzo, remained upstanding and productive members of society, thus providing much needed cover in the family for the prodigal son to pursue his academic passion. Finally, my caring and patient partner, Blanca, has constantly reminded me of the joy, liveliness and sparkle that machines cannot yet provide.

Index

Introduction: AI Concerns Us All

Roberto Musa G.

"I hold in awe both the discoveries of philosophy
and those who have made those discoveries;
and I thrill to claim what is, as it were, an inheritance from many predecessors.
Everything they collected, everything they labored over, was for me!
But let us do what a good head of household does: let us add to our endowment.
May it be a larger inheritance when it passes from me to posterity.
Much work remains to be done, and always will:
nothing prevents those born a thousand generations hence
from making their contribution."

Lucius Annaeus Seneca, 2005, p. 184

There is an old German saying: "*Es ist noch kein Meister vom Himmel gefallen*" (whose functional English rendering is the austere "practice makes perfect" while its literal meaning alludes to the unwillingness of the skies to rain ready-made experts down upon us). Gauss, Pascal and Newton have done their share to make us doubt the truism, but what can't be argued is that it certainly applies to the three essays that follow, which grew organically and laboriously out of my encounter with a domain that I perceive as gaining greater influence upon our collective future. Like a moth to a flame, I have been fascinated by the ever-increasing currency of ever-intelligent machines in our public discourse and our private lives, and I hope to have captured some of that wonder for the benefit of my readers in the pages that follow.

If media stories and statements by prominent intellectuals are to be taken as suitable indicators, there is a surge of renewed public interest in the development of Artificial Intelligence. The label has moved from being a distant and speculative buzzword into a household name. Should it surprise us that so many among our files walk the streets feeling they carry the seeds of an impending AI revolution already in their pockets, in the form of their "smart" (and who knows how long until those quotation marks of hesitation strike us as definitely unfair) assistants like Siri, Cortana, Alexa and (the less sexily named) Google Assistant? Grandmothers are worrying about it and for good reason. Elon Musk is worried too and so was the late physicist Stephen Hawking.

In a way, this should not be surprising for even if research in AI has had to face its fair share of budgetary winters throughout its rocky history—or perhaps rather, *sandy* history, as per AI's indebtedness to silicon—it is a topic that has always been able to grip the imagination. In the three essays that follow I hope to convince the reader that the roots of the fascination that AI arouses lie in the echoes it awakens in us. The booming rumble of such echoes harks back to the times in which the questions AI now brings to the forum were answered in the realms of the spiritual and the cosmological.

Layout

All three essays constitute different but interrelated approaches to an understanding of Artificial Intelligence in light of its history, its culture and its present condition. In these introductory words I will summarily give an overview of each of them, comment more generally on the paths I followed and the principles I adhered to in composing them (in connection to which I shall indulge a brief discussion of my tenets regarding scholarly work, which I believe to be warranted given the spirit of these writings), and highlight some key ideas of interest that may guide prospective readers to the essay or section most attuned to their present concerns. After this, I shall reflect on and try to make the case for what possible justification there is for someone who lacks formal training as a coder, computer scientist or engineer to write about a technical subject like AI, and even more importantly, why non specialists in that field would profit from reading about it too.

This assembly of writings deals with different aspects of Artificial Intelligence and the discourse that surrounds it. Each by itself and all three taken as a progression can be considered as addressing the past, present and future of Artificial Intelligence. While each of the three essays is self-contained and can be understood on its own, I believe there to be deep connections joining all three, whether stated or tacit. Therefore, in the general discussion that closes this volume I attempt to draw out at greater length what I see as the main threads running through all three seemingly idiosyncratic vantage points, tying them up with a topic where the ingenuity of the machine builder clashes with the sensitivity of the artist: creativity and whether machines are capable of developing it. Before I delve into what is to be found in each essay, allow me a few words on Artificial Intelligence in general. Scorn has been rightfully heaped on those prefaces that presume to explain succinctly but exhaustively what the whole book contains. What point then was it there to writing the book in the first place? Accordingly, the following is not an attempt at either a comprehensive definition or a thorough perspective on what

Artificial Intelligence is and entails (for that is what the thesis as a whole attempts to do), but rather, a bounded and functional point of departure.

I begin with the caveat that, while throughout all of these writings it may seem that I speak of Artificial Intelligence (or AI) as if it were a single and monolithic entity, or as if all of the researchers pursuing it were of one mind and in accord, in actuality this label must be understood as pointing to a diffuse field of seething complexity. In this light, a crucial aspect that must be taken into consideration is that far from being a distinctly delimited parcel of science, Artificial Intelligence is drastically characterized by its multidisciplinarity:

> AI is a multidisciplinary activity that involves specialists from several fields such as neuroscience, psychology, linguistics, logic, robotics, computer sciences, mathematics, social sciences, biology, philosophy and software engineering. And it covers a number of areas of interest, such as intelligence, the representation of knowledge, creativity, robotics, language translation, domotics, emotions, data mining, intentionality, consciousness and learning. (Vallverdú, 2011, p. 173)

Another aspect of AI's heterogeneity that dispels the hopes of easily pigeonholing a supposed monolith-like core of identity is a deep tension that traverses it since its very inception and that can be traced back all the way to its ancient past: the two interpretations of its main purpose reflected in its dual nature as a scientific and technological enterprise. In the words of Nils J. Nilsson, pioneer in AI research and its implementation in robotics:

> AI has as one of its long-term goals the development of machines that can do these things [perception, reasoning, learning, communicating, and acting in complex environments] as well as humans can, or possibly even better. Another goal of AI is to understand this kind of behavior whether it occurs in machines or in humans or other animals. Thus, AI has both engineering and scientific goals. (1998, p. 1)

That divide between—to employ Frederick Brooks's typology—the "scientist [who] builds in order to study" and "the engineer [who] studies in order to build" (1996, p. 62) is still alive and well today and we shall have occasion to observe its repercussions in the way AI is currently pursued. A similar tension that will also make frequent appearances is that identified by Jean-Pierre Dupuy (2000) between mechanizing the mind and humanizing the machine.

We begin with *Echoes of Myth and Magic in the Language of Artificial Intelligence*, an article whose theme was the initial reason behind my global interest in the topic. I noted from early on, that there were

very noticeable similarities between the language used in texts about AI and such writings as sacred texts, myths and folk tales. Some of these reverberations are so critical to the project's identity that they are repeatedly brought up by the AI research community, most notably the motif of trying to usurp God's powers by creating life, which is so clearly illustrated in the story of the Golem and in *Frankenstein*, its Victorian (pun intended) retelling. It is as if in every era we were telling ourselves, as humans, the same stories, but using slightly different language, one specially suited to the era we belong to:

> The makers of the golem, living in a God-dominated universe, never presumed to make anything more than a rough draft of a man. Frankenstein, living in the Age of Reason when there is no power higher than man, adds to the dreams of the magician the hubris of his age, in his aim to make a creature who will be so much an improvement on the original model that he will make man himself look like the rough draft. (Rowen, 1992, p. 175)

In this way, Biblical stories give way to the folk tale of the Golem which is succeeded by Mary Shelley's *Frankenstein*, which in turn ends up becoming the conversational part of a paper by Herbert Simon, Allen Newell, John McCarthy, Marvin Minsky or Norbert Wiener. Of course, at a high enough level of abstraction, any two things can be said to be identical. When viewed from the requisite distance, both Goethe's *Faust* and a washing machine's instruction manual are treatises on the overcoming of the innate limitations of men. This is why some scholars have sought to collapse all stories into a few overarching archetypes (e.g., Booker, 2004). However, we strongly believe that in likening these archetypal stories from different ages we are moved less by fancy than by their innate characteristics and historical relationships. Arturo Aldunate Phillips, Chilean science writer of poetic sensibilities and a thorough student of cybernetics (having been closely associated with Wiener) wrote poignantly of the dance between literature and science:

> So close lies science to dreams, reality to fantasy, that some of the preeminent scientists of our day must turn, in order to tell us what they glimpse from the horizon of their laboratories and what stirrings their findings spur, to the science fiction tale. "The Black Cloud", by Hoyle, the great British astronomer, and "The Tempter" by Norbert Wiener, the genius founder of cybernetics, illustrate this new and curious attitude of the men of science turned men of letters, using fable and fancy to pave the road towards truth. (Aldunate Phillips, 1964, p. 5, my translation)

But what may have felt like a novelty in the sixties, from where Aldunate Phillips writes, is avowedly par for the course today, especially when it comes to the realm of science fiction. We will delve deeply

into how science fiction nourishes mainstream AI research and is in turn influenced by it, and how it serves to shed light in the deep interrelationship between AI and religion:

> The blurring between science fiction and science fact is of considerable interest given that it shows the kinds of inspirations that scientists experience and also the way in which the future appears amenable to human intervention; more importantly, however, in Apocalyptic AI we see the sociocultural power of pop science. (Geraci, 2010, p. 41)

If there is one topic that keeps cropping up repeatedly in science fiction, it's that of the minds of intelligent machines and whether they are functionally and morally equivalent to ours. In *Computing Machinery and the Benefit of the Doubt* I consider this issue by taking a deep look at the Turing Test. Arguably, no other idea from the philosophy of Artificial Intelligence so captures the imagination of experts and lay audiences alike, and accordingly, it has been featured in a variegated array of pop culture products. This overexposure, however, has also led to a sort of guilt-by-association that paints as shallow or obvious one of the greatest insights to emerge from the philosophical underpinnings of Artificial Intelligence. For all its notoriety, the Turing Test runs the risk of becoming increasingly seen as a mere historic landmark in the evolution of thought on AI. However, I argue that in Turing's seminal paper itself lies a powerful mental tool that must be recovered and brought to attention, for it provides precious help in steering us into more empathetic and less solipsistic ways of being, and offers a roadmap to much-needed scientific integrity.

There is a line of criticisms leveled against Turing's test which can ultimately be pinpointed to a lack of the imagination needed to envision the full power of what the imitation game allows us to probe, which we do our best to argue against. Douglas Hofstadter offers a very illustrative analogy for the potential for unveiling mentality that Turing's protocol affords:

> Examining linguistic responses in novel ways is quite similar to examining the spectra of stars in novel ways (e.g., using new regions of the electromagnetic spectrum, higher degrees of resolution, time-correlation data from widely separated receivers, etc.), and thereby inferring detailed stellar mechanisms of ever subtler sorts, despite being hundreds of light-years from the star itself. (Hofstadter, 1995, p. 489)

In the third essay, *Gamifying Programs* we explore how games have shaped human evolution and AI development. By looking at the past history of many AI systems as game playing systems, we are offered a window into the painstaking and gradual steps that brought them forth, as well as an

opportunity to reflect on how much our own species' need to play influenced our endowing our mechanical progeny with a penchant for games. But the most interesting question is where we can anticipate this interrelated gamefulness to take us. What does that let us predict concerning our future interactions with machines? What kinds of games are we likely to play with them in the future? What games will they, instead, only play among themselves?

After a summary discussion around the concept of 'gamification' (a label that while being fairly recent, possesses underlying ideas that can be traced back many decades, at the very least to the beginnings of behaviorism), we analyze the ways in which our culture is becoming increasingly gamified day by day. Video games are becoming a cornerstone of the daily routines of many adults, especially males, and this trend can only be expected to keep rising. However, it is not just proper games per se which usher the gameful into daily life. Other spheres are made increasingly to resemble games, in the service of political or economical interests. Projecting the trajectory that we are currently embarked upon, we can predict that virtual reality will only increase its encroachment upon the daily lives of ordinary citizens. One of the traditional criteria that defined play was its closed nature, that is, its standing outside of normal life:

> A characteristic of play, in fact, is that it creates no wealth or goods, thus differing from work or art. At the end of the game, all can and must start over again at the same point. Nothing has been harvested or manufactured, no masterpiece has been created, no capital has accrued. (Caillois, 2001, p. 5)

But while play in itself may not be able to create longstanding wealth that outlasts its duration, the larger contexts in which it is embedded well might, as is attested by anyone who has paid the admission ticket to a professional sporting event. The current cutting-edge developments in social video games intend to profit from the wealth that can be created as a consequence of games in a far more concerted and thorough way. A telling trait of our times is that the spaces of work and play are no longer as distinct as they once were. As was put forth by game evangelist Jane McGonigal, "with gameful design, we are intentionally stepping outside the magic circle of play—or at the very least, fusing the magic circle with the ordinary world in ways that seek to change it" (McGonigal, 2014, p. 655).

Finally, before moving on to two further reflections, allow me to introduce a stylistic convention. Insofar as they are thematically interconnected, the five pieces of this thesis often refer to one another. For convenience, and as opposed to the formal requirements of the journals to which the three main articles were submitted, I adopt the following naming strategy throughout to refer to each of them:

[INTRO]: "Introduction: AI Concerns Us All"
[MYTHS]: "Echoes of Myth and Magic in the Language of Artificial Intelligence"
[TESTS]: "Computing Machinery and the Benefit of the Doubt"
[GAMES]: "Gamifying Programs"
[OUTRO]: "General Discussion: Creators, Creatures and Creativity"

The Scholar as Forager

In his analysis of how the personal perspectives and experiences of economists impact their scientific activity, David Carré (2017) justly calls attention to the words of John Maynard Keynes: "Practical men, who believe themselves to be quite exempt from any intellectual influences, are usually the slaves of some defunct economist" (2007, p. 383). It is enlightening to compare these remarks on the lasting afterglow attained by a set of long-delivered ideas to a very similar formulation concerning the realm of AI. John McCarthy, who christened the field as we now know it, stated that "AI research not based on stated philosophical presuppositions usually turns out to be based on unstated philosophical presuppositions. These are often so wrong as to interfere with developing intelligent systems" (McCarthy, 1999, p. 72).

Therefore, while AI is heralded as one of the newest trends in human activity (and in a sense, it certainly is) we gain much from studying its long and rich history. In [MYTHS] we will in fact see how the history of AI can be considered a microgenetic recapitulation of longstanding debates in philosophy. No matter where this inquiry takes us, it is worth bearing in mind that no history of AI is divorced from a reflection about ourselves and that all that is written about our intelligent machines must be understood in this connection. Several challenges spring to mind in light of the aims of the present research. Chief among those is the responsibility of portraying the views of different researchers in as fair and accurate a light as possible. Thinkers, writers and theoreticians are unique individuals and hold nuanced and variegated assemblies of opinions and perspectives. The slapping of "-isms" may be a convenient way to group them, but necessarily compromises an enriched understanding of what each is advancing. In a way, exemplifying just to what extent this occurs is one of the core issues that these essays attempt to deal with.

The pieces in this thesis are differently shaped windows from which to look at a unitary phenomenon. In [TESTS] I argue that striving for exhaustivity in this day and age is a doomed errand. However, my main goal is to make my reader profit from my hours of reading among the sources and curating the

choicest of bits. Being possessed of too trigger-happy a disposition to relay quotations from venerable authors is considered a shameful trait by some of them (like Emerson and Seneca, whose superb remarks on the matter decorum forbids me from quoting verbatim so as not to infringe upon their wishes). But I strongly believe that there is real merit in gathering them and offering them up again. If we grant, as we should, that the classics are those works which are the least suited for informational compression, then, still, a choice passage comes far closer to giving us a true flavor for the whole text, as an organic microcosmic embodiment of its larger structure, than does a summary that needs must traverse the diffraction of the summarizer's mind. Seen in this light, the explorers that come before us and who unearth such precious gems are wholly deserving of their share of the credit in making their finds more easily available to the community.

Verónica Watt (2011), in her investigation of the functions that epigraphs carry out, quotes John Barth on the peril of being outclassed outright by the words we have chosen to present before our own: "to preface your text with an epigraph from a superior author in the same genre is to remind the reader that he might better spend his time with that author than with you" (1997, p. 8).[1] I see the logic but am nevertheless troubled, as naïve as it may sound. Given that we all (at least up to this point in the progression of medical technology) lead mortal lives and our allotted time for reading is finite, I would indeed much prefer for readers to spend theirs with texts better than my own. I have made thus my best effort to render as transparent as I could the path that I followed, so that my excursions through the literature could be easily retraced and expanded upon by anyone who should be so inclined.

Perhaps this is why Charles Babbage, august grandfather[2] of computing and a character we will have ample occasion to get to know better in the course of the three main pieces, seems to have instinctively abided by this avoidance of uneasy comparisons, therefore choosing as epigraph for his autobiographical memoirs the laments of a fictional oyster of philosophical temperament, gravely misunderstood by the world in its humble ambitions to "make out a complete system of the universe, including and comprehending the origin, causes, consequences, and termination of all things" (1864, p. 5). I feel emboldened in the course of action I have set myself upon for being in the good company of many of the voices that shall accompany us throughout our explorations of the cultural history of AI. Here is, for instance, one such prefatory caveat:

> I have borrowed freely from the writings of several authors, and I have tried to acknowledge the source when I could remember it. Dr. Johnson thought that a man could turn over a library to make a book; even in these degenerate days it is accounted plagiarism to copy from one author. Academic tradition

> now accepts the convention that to crib from more than two counts as research […] I have tried not to follow the precedent established by the author of a famous treatise on Chinese Metaphysics, who, as the reader will doubtless recall, read the articles in the encyclopedia on China and on Metaphysics, and "combined the two." Much of this book derives from the work of those prolific authors "Anon" and "'Ibid" who have done so much to put our English platitudes on a sound literary basis. (Bowden, 1953, p. xii)

This asserts B.V. Bowden nonchalantly, with the distinctive bonhomie that pervades his *Faster than Thought*, a work that could rightly be considered the first book in the field of computer science (and in which with complete freedom, for instance, he intersperses a digression on the merits of Arab horses into his archeology of computing). This book houses Turing's most detailed writings on chess, which we will examine in depth in [GAMES] as well as the more puzzling preliminary remark that "machines are much less co-operative than human beings in telepathic experiments" (Bowden, 1953, p. 287), something that we shall explore further in [TESTS].

Johann Huizinga, a tutelary figure whenever one embarks in writing about games (and who shall thus deservedly provide us with a starting point for our reflections in [GAMES]) prefaced his magisterial study of them with the following words, that testified to the unescapable truth that even scholars of the highest caliber cannot escape from having to tread upon terrain that has already been fruitfully foraged in by their precursors:

> The reader of these pages should not look for detailed documentation of every word. In treating of the general problems of culture one is constantly obliged to undertake predatory incursions into provinces not sufficiently explored by the raider himself. To fill in all the gaps in my knowledge beforehand was out of the question for me. I had to write now, or not at all. And I wanted to write. (1980, p. x)

And in the words of C.G. Darwin, Turing's superior at the National Physical Laboratory, and arguably the man responsible for delaying by a couple of years the theorization regarding Artificial Intelligence (about whom more also in [GAMES]) began his own foray into what the future of humanity held in store by exculpating himself from not being as thorough as he would have wished in the relaying of his sources, but hoping the world would both forgive him for it and correct what mistakes he might thus introduce:

> The spirit of criticism is much commoner in the world than the spirit of invention, and progress has often been delayed by authors, who have refused to publish their conclusions until they could feel they had reached a pitch of certainty that was in fact unattainable. Progress in knowledge is more rapidly made by taking the chance of a certain number of errors, since both friends and enemies are only too pleased to exert their critical faculties in pointing out the errors; so they are soon corrected, and little harm is done.[3] (Darwin, 1952, p. 8)

In *An Essay Concerning Human Understanding*, John Locke famously said that he "that has but ever so little examined the citations of writers, cannot doubt how little credit the quotations deserve, where the originals are wanting; and consequently how much less quotations of quotations can be relied on when the originals are wanting" (1690/1999, p. 660). It constitutes a terrible irony, therefore, that nowadays a large majority of people come across those words of his in the pages of *Bartlett's Familiar Quotations* or in the citations of other writers, than nearly by the closing of the *Essay* itself. And truth be told, even after examining with care the section from where Locke's words emerge, we find that the quote-worthy excerpt quite excellently conveys the intended meaning of the whole.

Whomever sets foot on the wilderness of the originals should aim to carry back to the tribe the treasures therein found. And like strange and beautiful animals, while it is true that their strong willed and supple forms are best appreciated in their natural environs, the fact of the matter is that were it not for zoos, most of them would never be seen at all. And more than one species has been saved from the brink of extinction for its having been thus preserved in the chronicles of those who came after. Related to the above, it must be specified however that the value of curation lies not only in boosting the extant stigmergic signals of the oft-quoted when warranted, but even more importantly, in rescuing the valuable but obscure (more on the repercussions of this idea in [OUTRO]).

Of course, this shuffling of the past, along with its occasionally felicitous rediscoveries will offer its fair share of dead ends and disappointments, such as those crowning the hours spent, upon having learned of its existence, tracking down a letter Charles Babbage sent to Francis Galton. Given the stature of both thinkers, I relished in imagining what precious insight these two minds might have shared on the descent of men and the descent of their factories, only to discover that the letter was nothing else than a simple note of thanks (and a nearly illegible one at that) for the gift of a book on the topic of Australia.

Why Not Just Leave AI to the Specialists?

There have been numerous serious attempts to understand AI and its cultural impact. More specialized treatises exist for those seeking to delve deeper (e.g., Boden, 2006; Ekbia, 2008; Hofstadter, 1979/2000; McCorduck, 2004) and I will go out of my way to promote them when due. So what possible justification could there be for adding one more into the mix? And furthermore, why should this be of special concern to psychologists? The answer lies in that whether we know how to program or not, all of our lives are already being affected by the consequences of Artificial Intelligence, and will only be more so affected in the future. Therefore, we have a right to ponder deeply the directions in which we are collectively embarking (or, perhaps, being led) and raise our voices adding to the discussion. After all, while scientists and engineers might be very pure in their devotion to the pursuit of their goals, the public at large has an inkling that this lust for knowledge does not always work in its best interest: "The suspicion that the scientist is not quite sincere in professing that his purpose is purely mechanical and illustrative goes a long way back. The notion of magic is deep-rooted." (Walter, 1961, p. 104)

There is a canonical line from the movie *Jurassic Park* that is frequently invoked whenever scientists create havoc in their zeal to advance their learning or instantiate their theories: "Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should". It is the equivalent of that oft-given reason for climbing Mount Everest: "Because it was there." The whole essence of *Frankenstein* is centered around this dilemma. In the final section of [MYTHS] we address these concerns more directly, but it's worth bringing them up here as well, as they provide a rationale for our approach. James Barrat, a documentary filmmaker turned AI doomsayer encapsulates the idea vividly:

> I think there's a high chance of painful mistakes on the way to AGI [Artificial General Intelligence], as well as when scientists actually achieve it. As I'll propose ahead, we'll suffer the repercussions long before we've had a chance to learn about them […] As for the likelihood of our survival—I hope I've made it pretty clear that I find it doubtful. But it might surprise you to know my chief issue with AI research isn't even that. It's that so few people understand that there are any risks at all involved along AI's developmental path. People who may soon suffer from bad AI outcomes deserve to know what a relatively few scientists are getting us all into. (Barrat, 2013, p. 117)

Even if we don't buy into as gloomy a scenario as the one predicted by Barrat, still, the increasing impingement of AI and algorithms into our lives are cause for concern and should move us to reflect on them in order to take action. We see this encroachment patently in the way that the bureaucratic red tape that we must outmaneuver gets increasingly algorithmized: "As individuals and as a society, we increasingly depend on artificial intelligence algorithms we don't understand. Their workings, and the motivations and intentions that shape their workings, are hidden from us. That creates an imbalance of power" (Carr, 2015 p. 38). We should also have a say on the way that AI will impact our economic future: "Policy decisions about how to share society's growing wealth will impact everybody, so the conversation about what sort of future economy to build should include everyone, not merely AI researchers, roboticists and economists" (Tegmark, 2017, p. 122).

After all, among the many difficulties in accurately assessing how smart our current algorithms really are, market forces stand out. For just as in the fifties the capacities of computers were downplayed so as not to scare consumers away from adopting an unfamiliar technology, thus being bestowed the label of quick morons, that many still associate with them to this day (McCorduck, 2004), computers being no longer unfamiliar, we see the opposite trend, since now that labels like *AI* and *Deep Learning* have become sexy it seems that many companies are eagerly attempting to stamp whatever they are working in with them (du Sautoy, 2019).

One of the biggest risks before us is one that has been constantly heralded by previous critics of the rising automation that as a society we are facing. Namely, that in increasing the luxuries we have access to by means of industrial mechanization, we will become so reliant on these technologies that we will end up weakened, and completely dependent upon them. In connection to this idea I will present some of Theodore Kaczynski's prescient words of caution. I have chosen to discuss his thought, also, because he illustrates something that I argue passionately in [TESTS]: that ideas must be judged on their own merits, regardless of what may be ascribed to those who espouse them.

The future of artificial intelligence is inextricably entwined with global trends such as geopolitical power contests and the future of capitalism. After all, much if not most of the research budget devoted to AI has come from military funding, seeking technological innovations with practical applications in the battlefield (Geraci, 2010; Barrat, 2013; Tegmark, 2017). In [GAMES] we broach the topic of how this technomilitarization will blend with more innocuous technologies to produce ever greater surveillance (benefitting perhaps Foucauldian scholars who might see a well-merited

increase in their research budgets). If we fail to seek to influence the directions these technologies will take, we will be surrendering a sizable portion of our future:

> We incur another risk whenever we try to escape the responsibility of understanding how our wishes will be realized. It is always dangerous to leave much choice of means to any servants we may choose—no matter whether we program them or not. For, the larger the range of choice of methods they may use, to gain for us the ends we think we seek, the more we expose ourselves to possible accidents. We may not realize, perhaps until it is too late to turn back, that our goals were misinterpreted, perhaps even maliciously, as in such classic tales of fate as Faust, the Sorcerer's Apprentice, or The Monkey's Paw. (Minsky, 1984, p. 69)

That is, there are no guarantees that intelligent systems will care about any of the things we care about. That is why it is so critical to get their programming just right, because if we fail to align their goals to ours, the consequences we will have to face may be life-ending (see [MYTHS]). But this should not lead us to despair. While blind optimism is leaving the front door of our mind and of our future open to all manner of risks, an utter lack of all hope whatsoever is but a latent suicide apathetically waiting for a thin spur of motor activity to bring it to fruition.

Here I cannot but quote an aphorism that has had a lasting influence on my thinking: Niels Bohr's provoking remark to the effect that the opposite of a small truth is a lie, but the opposite of a big truth is another big truth. The groupie and the luddite are alike ill-suited for the task we must set before ourselves; we should not side with Cassandra but neither with Pollyanna. As in most realms of human thought, a partisan point of view is unlikely to be of much use when considering a topic of such nuance and subtlety, and furthermore, one that has engaged some of the finest intellects the past and present centuries have produced, and had them, often, at odds with one another. With some trepidation, for fear of giving off undertones of being possessed by a distasteful noncommittal theoretical centrism and merely having imported into the discussion of AI the school of political analysis of the have-it-and-eat-it-too cakeism, I must declare that we are enriched when we approach two contending views on the same matter with an open mind.

After all, when even so subtle and profound a thinker as Jean-Pierre Dupuy accuses the modern golem builders of being as empty as their creations supposedly will be, we realize that the debate is in desperate want of sensible middle ground, where rather than lashing out in fierce attacks, we can

understand that both those who seek to usher in the age of intelligent machines and those who seek to repel it are driven by deeply human motivations:

> The most perfect simulation still fails to capture something, and it is this something that is the essence of love—this poor word that says everything and explains nothing. I very much fear that the spontaneous ontology of those who wish to set themselves up as the makers or re-creators of the world know nothing of the beings who inhabit it, only lists of characteristics. (Dupuy, 2000, p. xx)

This is nonsense. Well-meant nonsense but nonsense nonetheless, and I hope I have been able in the essays that follow to present a picture of AI researchers that does more justice to them than such highbrow accusations of unworldliness. In that spirit, I feel I take up the same approach of Pamela McCorduck, who addressed those very criticisms, here offered by apostate Joseph Weizenbaum once embarked on his denouncing of the field. Weizenbaum's objection being that "AI workers are single-minded about the model as the sole means of encompassing the human experience" (cited in McCorduck, 2004, p. 377), to which she sagely counters as follows:

> That isn't my impression. Their writings are enthusiastic but tentative: they think they're onto a good thing, maybe one of the best so far, as an approach to explaining human cognition. But they read novels and poetry, compose and play music, see movies and write stories, and make the same noises about the value of doing those things as the rest of us. (McCorduck, 2004, p. 377)

"The intuitions of a lover are not always to be trusted; but neither are those of the loveless" said Joseph Wood Krutch (1959, p. xi) when defending his right as a mere lover of nature to speak on scientific matters. And we must avow that AI researchers do love their machines, even if at times, they end up recoiling in despair at what they have created, as Weizenbaum certainly did with ELIZA and on the basis of which B. Jack Copeland (1993) compared him to a young Victor Frankenstein, fleeing from his creation.

References

Aldunate Phillips, A. (1964). *Los robots no tienen a Dios en el corazón*. Santiago, Chile: Editorial Andrés Bello.

Babbage, C. (1864). *Passages from the Life of a Philosopher*. London: Longman, Green.

Barrat, J. (2013). *Our Final Invention: Artificial Intelligence and the End of the Human Era*. Chicago: Thomas Dunne Books.

Barth, J. (1997). *The Friday Book: Essays and Other Nonfiction*. Baltimore: Johns Hopkins University Press.

Booker, C. (2004). *The Seven Basic Plots: Why We Tell Stories*. London: Bloomsbury Continuum.

Boden, M.A. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford: Clarendon Press.

Bowden, B.V. (1953). *Faster than Thought: A Symposium on Digital Computing Machines*. London: Sir Isaac Pitman & Sons.

Brooks, F.P. (1996). The computer scientist as toolsmith. *Communications of the ACM*, 39(3). pp. 61–68.

Caillois, R. (2001). *Man, Play and Games*. Urbana and Chicago: University of Illinois Press.

Carr, N. (2015). The Control Crisis. In J. Brockman (Ed.), *What to Think About Machines That Think*. NY: HarperCollins. pp. 59-61.

Carré, D. (2017). *Towards a Cultural Psychology of Science: Economics and Economists in Contemporary Chile*. (Doctoral dissertation). Aalborg University, Denmark.

Copeland, B.J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

Darwin, C.G. (1952). *The Next Million Years*. London: Rupert-Hart Davies.

du Sautoy, M. (2019). *The Creativity Code: How AI is Learning to Write, Paint and Think*. London: 4th Estate.

Dupuy, J-P. (2000). *On The Origins Of Cognitive Science: The Mechanization Of The Mind*. Princeton: Princeton University Press.

Ekbia, H. (2008). *Artificial Dreams: The Quest for Non-Biological Intelligence*. Cambridge: Cambridge University Press.

Geraci, R.M. (2010). *Apocalyptic AI: Visions of heaven in robotics, artificial intelligence, and virtual reality*. Oxford: Oxford University Press.

Hofstadter, D. (1995). On Computers, Creativity, Credit, Brain Mechanisms, and the Turing Test. In D. Hofstadter and the Fluid Analogies Research Group (Eds.) *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. NY: Basic Books.

Hofstadter, D. (2000). *Gödel, Escher, Bach: An Eternal Golden Braid*. London: Penguin. (Original work published 1979).

Huizinga, J. (1980). *Homo Ludens: A Study of the Play Element in Culture*. London: Routledge.

Keynes, J.M. (2007). *The General Theory of Employment, Interest and Money*. London, UK: Palgrave Macmillan.

Krutch, J.W. (1959). *The Great Chain of Life*. London: The Country Book Club.

Locke, J. (1999). *An Essay Concerning Human Understanding*. Pennsylvania: Pennsylvania State University. (Original work published 1690).

McCarthy, J. (1999). Philosophical and Scientific Presuppositions of Logical AI. In H. Levesque & F. Pirri (Eds.) *Logical Foundations for Cognitive Agents*. Berlin: Springer, pp. 72–78.

McCorduck, P. (2004). Foreword to *Machines Who Think*. Natick, Massachussets: A K Peters.

McGonigal, J. (2014). I'm Not Playful, I'm Gameful. In S.P. Walz & S. Deterding (Eds.) *The Gameful World: Approaches, Issues, Applications*. Cambridge, MA: MIT Press, pp 653–657.

Minsky, M. (1984). Afterword to 'True Names'. In V. Vinge *True Names*. NY: Bluejay Books.

Nilsson, N.J. (1998). *Artificial Intelligence: A New Synthesis*. San Francisco, CA: Morgan Kaufmann Publishers.

Rowen, N. (1992). The Making of Frankenstein's Monster: Post-Golem, Pre-Robot. In E. Ruddick (Ed.) *State of the Fantastic: Studies in the Theory and Practice of Fantastic Literature and Film*. Westport: Greenwood. pp. 169-77.

Seneca, L.A. (2015). *Letters on Ethics to Lucilius*. [Translated and edited by M. Graver and A. A. Long]. Chicago: University of Chicago Press.

Smith, J. & Smith, H. (1879). *Rejected Addresses: Or, The New Theatrum Poetarum*. London: John Murray.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. NY: Alfred A. Knopf.

Vallverdú, J. (2011). The Eastern Construction of the Artificial Mind. *Enrahonar: quaderns de filosofia*, (47), 171–185.

Walter, W.G. (1961). *The Living Brain*. UK: Penguin Books.

Watt, V. (2011). *Funciones epigráficas en novelas de autores norteamericanos contemporáneos*. (Masters Dissertation). Universidad Diego Portales.

Echoes of Myth and Magic in the Language of Artificial Intelligence

Roberto Musa G.

"We have lived so long with the conviction that robots are possible,
even just around the corner,
that we can't help hastening their arrival
with magic incantations."

Drew McDermott, 1981, p. 145

Abstract

To a greater extent than in other technical domains, research and progress in Artificial Intelligence has always been entwined with the fictional. Its language echoes strongly with other forms of cultural narratives, such as fairytales, myth and religion. In this essay we present varied examples that illustrate how these analogies have guided not only readings of the AI enterprise by commentators outside the community but also inspired AI researchers themselves. Owing to their influence, we pay particular attention to the similarities between religious language and the way in which the potential advent of greater than human intelligence is presented contemporarily. We then move on to the role that fiction, science fiction most of all, has historically played and is still playing in the discussion of AI by influencing researchers and the public, shifting the weights of different scenarios in our collectively perceived probability space. We sum up by arguing that the lore surrounding AI research, ancient and modern, points to the ancestral and shared human motivations that drive researchers in their pursuit and fascinate humanity at large. These points of narrative entanglement where AI meets the wider culture should serve to amplify the call to engage ourselves with the discussion of the potential destination of this technology.

Keywords: Artificial Intelligence, religion, science fiction, existential risk, philosophy of science, technological singularity

---

Questions about AI inextricably lead to wondering what it means to be human and where exactly the boundaries lie of that which defines us. After all, the computer is "the most complex technology ever devised by man, and we hold it up as a mirror to our own souls"[4] (Fellows, 1995, p. 85). When considering what could have possibly motivated the participants and organizers of the Dartmouth Conference (where first the field got its moniker) to "devote their professional lives […] to building machines either to mimic the human brain or to behave intelligently, by hook or by crook" (McCorduck, 1979, p. 134), Pamela McCorduck, celebrated chronicler of the dawn of AI, reports several alternatives "offered by armchair psychologists" (p. 134), counting such variegated possibilities as "the desire to be as gods", being able to "have offspring without the help or interference of a woman", the Freudians' suggestions of "a yearning to desexualize or cleanse procreation, counter-pointed by the Oedipal drama" or "an urge to divide the self, to make a doppelganger that would carry away the evil in one's soul, leaving of course the residue of good." In the end she supposes that, as so often is the case, the purloined letter explanation lying in plain sight is the most apt:

> But perhaps the main reason is also the most obvious one. To know intelligence well enough to be able to build a working model of it is surely one of the most intellectually exciting and spiritually challenging problems of the human race. To do so is to know ourselves as we've always yearned to, to make us a part of nature instead of apart from it, in Herbert Simon's felicitous phrase. Such knowledge implies a solution of the mind-body problem, which has eluded the most intense human efforts for over two thousand years. And such a model promises to be an extension of those human capacities we value most, our identifying properties, which we sum up as our intelligence or our reason; the thinking machine would amplify these qualities as other machines have amplified the other capacities of our body. (McCorduck, 1979, p. 135)

We may see in AI's research program a trace of the same spirit that imbued Vico's principle of *Verum et factum convertuntur*: we can only truly know that which we have created ourselves, and so it is that man may understand culture, but nature is accessible only to God (Vico, 1948). Therefore, the royal road to understanding the mind would be to create one. This is strikingly similar to the sentiment expressed by physicist David Deutsch, in elaborating on the merits of Turing's test: "I have settled on a simple test for judging claims, including Dennett's, to have explained the nature of consciousness (or any other computational task): if you can't program it, you haven't understood it" (Deutsch, 2011, p. 132). Even should those attempts ultimately fail, we would have gained precious insight into our very nature,

as even one of the harshest critics of the AI enterprise will readily attest: "What we learn about the limits of intelligence in computers will tell us something about the character and extent of human intelligence" (Dreyfus, 1992, p. 79).

Literature—understood in the broadest of senses— has devoted relentless attention to these questions of the limits between mankind and its creations and has also heaped precious insight upon them. In following the traces of what the AI community has harvested out of that literary treasure trove, we will now focus on myths and religious writings and then move on to science fiction, both classical and contemporary. Taken together, such a corpus could be considered the collective *Bildungsroman* of our species or, perhaps, of a new one.

The Sorcerer's Apprentice

The vast storehouse of old myths is rich in stories of those who met their demise by trying to emulate the gods. It was their pride in trying to obtain the Creator's power or acquire abilities that would make them superior to their peers that doomed Icarus and Daedalus or the makers of the Tower of Babel. Phaethon lusts for a power that he cannot contain and is struck down by Zeus while in his unruly handling of the chariot of the sun, to prevent him from visiting destruction upon the world. However, no offense committed by the brethren of Prometheus—who, with his mythical stealing of the fire and ushering in of mankind's technological age, deserves to be counted as the spiritual forerunner of the lot—is as egregious as the attempt to usurp God's most holy attribute and create life.

We see that quite distinctly in the story of Doctor Victor Frankenstein, whom Mary Shelley (1818) deservedly dubbed "the modern Prometheus". The novel—far richer than the social representation evoked by the name 'Frankenstein' in the minds of those who only know it through the movies or its even more diluted trickling down into pop culture—has, like all classics, much to teach us. In [TESTS] we explore what lessons can be gleaned from it that would help us understand the Turing Test as an analytic device that aids us in navigating our relationship with our fellow beings with empathy and scientific integrity. Let us now turn to what it shows of the risks of AI in general.

Much like Faust, Victor Frankenstein has drunk from the vial of science to its dregs, and remains thirsty still. He seeks to surpass all his predecessors by attaining that which has been achieved solely by God—or, in a secular reading, that blind, idiot god, Evolution (Yudkowsky, 2007b)—and bestow inert matter with life. His actions find such dire consequences in the grim retribution of his creature,

that the template of the story has pervaded our thinking about robots to the extent that Isaac Asimov (a tutelary figure for many an AI researcher) "called the fear of humanlike machines the 'Frankenstein Complex.'" (Foerst, 2004, p. 31)

Frankenstein is itself a modern retelling of the ancient legend of the golem, in which the Rabbi of Prague creates a humanoid out of clay who is animated by inscribing on his forehead the name of God (or alternatively, in other versions, the Hebrew word *emet*, or 'truth'). The golem narrative and its derivatives have played an undeniably significant role in our shared cultural understanding of the AI enterprise. Comparisons between the golem and computers endowed with human-like cognition have been explicitly touched upon, not merely by contemporary literary theorists but also by distinguished pioneers from the field. Paramount among these is the founder of Cybernetics, Norbert Wiener[5] who, in his aptly titled book *God, Golem, Inc.*, made the connection quite explicitly: "The machine, as I have already said, is the modern counterpart of the Golem of the Rabbi of Prague." (1964, p. 95)

Mitchell Marcus, former chair of the Computer and Information Science Department at the University of Pennsylvania and a graduate from the MIT Artificial Intelligence Lab has explicitly drawn the comparison as well. As reports George Johnson in his article for the New York Times on the "Science and the Spiritual Quest" conference organized by the Templeton Foundation in 1998 to promote dialogue between science and religion, Marcus gave a speech therein stating that:

> the craft of artificial intelligence—designing thinking computers—is a modern realization of the school of Jewish mysticism based on the Kabala. According to this ancient teaching, it is not quarks and leptons but the first 10 numbers and the 22 letters of the Hebrew alphabet that are the true fundamental particles: the elements of the divine utterance that gave rise to creation. "Computer scientists," he declared, "are the Kabalists of today." The ancient rabbis are said to have used magical incantations to create beings called golems. The programmers create their simulated creatures with incantations of computer code. (Johnson, 1998, ¶ 28)

In Brian Lancaster's (2007) book on the Kabbalah, there is reported a more direct link still between Artificial Intelligence and mysticism which enhances the thematic link between both domains. Even if highly dubious, verging on the domain of the apocryphal anecdote or falling under suspicion of being no more than a practical joke, the story deserves to be shared. Marcus recounts that during his time at MIT he learnt of an astonishing story involving three Jewish AI pioneers from MIT; Joel Moses, Gerry Sussman and their famed teacher, Marvin Minsky. Moses told that on the occasion of

his bar mitzvah his grandfather called him apart to tell him that he was a descendent of the actual Rabbi of Prague who had created the original golem, and furthermore that the golem had not been destroyed, as the legend claimed, but was actually dormant in suspended animation. He then proceeded to bestow upon him the secret spell that could awaken the golem, entrusting him to transmit it in turn to future generations. After hearing this, Sussman was speechless. He had been told the exact same story by his own grandfather on his bar mitzvah. Supposedly, each of them then proceeded to go to a corner of the room and write down the spell independently. When they compared both spells, these turned out to be equal. Suddenly, Minsky came out of his office and seeing the students in such a state of shock he asked what was going on. After hearing the story, he said it was utter nonsense, for he too had heard that from his own grandfather on his own bar mitzvah, but had not believed it for a second (Lancaster, 2007, p. 187).

Theologian Anne Foerst, who was closely connected to the AI community at MIT, where she founded and directed the *God and Computers* project, relays a very similar version of this story in her book *God in the Machine: What Robots Teach Us about Humanity and God* (2004, p. 39). Further supporting evidence by a contemporary of those involved lends added credence to the account:

> Curiously enough, several present-day researchers in artificial intelligence have told me that they grew up with a family tradition that they are descendants of Rabbi Loew, though they doubt this belief has had much influence. Among them are Marvin Minsky and Joel Moses of M.I.T. Further, Moses tells me that a number of other American scientists have considered themselves to be descendants of Rabbi Loew, including John von Neumann, the computer pioneer, and Norbert Wiener, who coined the term cybernetics. (McCorduck, 1979, p. 13)

Interestingly, Lancaster adds that not only is the narrative of AI influenced by the story of the golem, but in turn, that the early roots of the golem story contained neither the element of the golem serving as a slave of its human masters nor the danger of it growing out of control and threatening their lives. That idea would have arisen subsequently from the influence caused by shifts in the social and cultural outlook regarding modern science (Lancaster, 1997). And in Foerst's (2004) reading of the golem stories, their creation is not so much an act of hubris as one of godly worship, something coherent with Gershom Scholem's claim that "traditionally, golem-making had a psychic rather than practical purpose" (Scholem cited in Comrada, 1995, p. 245).

The fear of the golem is not merely allegorical but reflects the general fear of the machine, most particularly those ominous machines which, on the one hand, are like us but not quite like us while, on the other, they could excel over us so easily as to end up entirely replacing us. Samuel Butler, of *Erewhon* fame and Lamarckian leanings, dealt in fiction with a world in which that danger came to pass. But he also considered it, way back in 1863, a very real possibility to be taken seriously:

> We refer to the question: What sort of creature man's next successor in the supremacy of the earth is likely to be. We have often heard this debated; but it appears to us that we are ourselves creating our own successors; we are daily adding to the beauty and delicacy of their physical organisation; we are daily giving them greater power and supplying by all sorts of ingenious contrivances that self-regulating, self-acting power which will be to them what intellect has been to the human race. In the course of ages we shall find ourselves the inferior race. (Butler, 1863, ¶ 5)[6]

The references to the golem mentioned so far contain a detail that must be highlighted for it will be of great importance when we move on to the discussion of perceived AI risk: the isometric relationship between magic and AI is deeply reflected in the symmetry between coding and knowing the sacred words of a spell. The imperative of utmost formulaic accuracy passed from Rabbi Löw to his alleged spiritual descendants has deep historical roots and tenacious conceptual tendrils:

> Coding is the primary tool of modern scientists and gamers who try to make digital artifacts, and coded incantations that derive from occult knowledge are the first methods that Renaissance scientists resorted to when trying to create and control their artificial servants and intelligent artifacts. (LaGrandeur, 2003, p. 1)

> More particular parallels exist between the metaphors that are integral to the cultures of computer scientists and early modern occult scientists. Both depend on understanding a secret language, both rely on personal illumination available in books, and both belong to societies of initiates which are seen by the rest of society as wielding their esoteric knowledge to do wonders (sometimes dubious wonders). (LaGrandeur, 2003, p. 2)

> Modern *computer* wizards use the information inherent in symbolic, programming language—their own form of incantations—to program systems that embody impressive aspects of human cognitive capabilities and, often, formidable physical power, such as is built into robots and Artificial Intelligence. (LaGrandeur, 2003, p. 4, emphasis in the original)

Having said all this it is, nevertheless, advisable to take a sobering step back so as not to be completely swept away by the force of the metaphor (and just how difficult it is to brace ourselves against the rushing tide of an aesthetically pleasing analogy!), in order to point out a noteworthy shortcoming. For all that talk of enchantments and incantations, of spells and chants of resounding magic, the similitude between the *language* of magic and myth and the *language* of AI, refers almost exclusively to the fossilized dimension of language as captured in the written word to the exclusion of actual living utterances, reducing language to nothing more than logic and losing what is central to human speech. The readiness of this intuitive and subtle interpretation is evidence of the primacy of the written versus the spoken word, which has unfortunately become the prevailing metaphor in language research (Ingold, 2017; Cornejo & Musa, 2017).

This observation also enriches the context for understanding Wiener's apprehensions regarding the inherent risks of instructions delivered to automata, to which we will now turn. Expressive and affective elements of speech being absent, the likelihood of misinterpretations regarding what is actually meant and wanted by the issuer of the command increases pointedly. Now, when it comes to the perils entailed by Artificial Intelligence and those of magic the parallels run deeper still. Wiener's words on the implications of the eventual rise of intelligent machines, which already in 1964 he envisioned as plausible, are so prescient and to the point as to deserve extensive reproduction:

> I am most familiar with gadget worshippers in my own world, with its slogans of free enterprise and the profit-motive economy. […] Power and the search for power are unfortunately realities that can assume many garbs. Of the devoted priests of power, there are many who regard with impatience the limitations of mankind, and in particular the limitation consisting in man's undependability and unpredictability. You may know a mastermind of this type by the subordinates he chooses. They are meek, self-effacing, and wholly at his disposal […] Once such a master becomes aware that some of the supposedly human functions of his slaves may be transferred to machines, he is delighted. At last he has found the new subordinate—efficient, subservient, dependable in his action, never talking back, swift, and not demanding a single thought of personal consideration. […] *This type of mastermind is the mind of the sorcerer in the full sense of the word. To this sort of sorcerer, not only the doctrines of the Church give a warning but the accumulated common sense of humanity, as accumulated in legends, in myths, and in the writings of the conscious literary man.* All of these insist that not only is sorcery a sin leading to Hell but it is a personal peril in this life. *It is a two-edged sword, and sooner or later it will cut you deep.* (Wiener, 1964, p. 53, emphases added)

Wiener is explicitly pointing at human hubris and ambition as the cause of the tragedy that could unfold, but he also posits a specific key point that explains precisely what it is that could go so wrong that the tragedy should occur:

> The theme of all these tales [he is alluding not only to the golem stories but also to *The Monkey's Paw*, *The Sorcerer's Apprentice* and *The Fisherman and the Jinni*] is the danger of magic. This seems to lie in the fact that the operation of magic is singularly literal-minded, and that if it grants you anything at all it grants what you ask for, not what you should have asked for or what you intend. If you ask for £200, and do not express the condition that you do not wish it at the cost of the life of your son, £200 you will get, whether your son lives or dies. The magic of automation, and in particular the magic of an automatization in which the devices learn, may be expected to be similarly literal-minded. If you are playing a game according to certain rules and set the playing-machine to play for victory, you will get victory if you get anything at all, and the machine will not pay the slightest attention to any consideration except victory according to the rules. If you are playing a war game with a certain conventional interpretation of victory, victory will be the goal at any cost, even that of the extermination of your own side, unless this condition of survival is explicitly contained in the definition of victory according to which you program the machine. (Wiener, 1964, p. 62)[7]

As we can see even more explicitly in his treatment of the Goethe-written and Disney-popularized tale of *The Sorcerer's Apprentice*, Wiener is emphasizing the warning that when you are working with spells, you incur in great risk when you do not master the precise words of the incantation, when you make the simplest of mistakes in the code. As Stephen Clark (1995) pointed out, Rudyard Kipling had earlier issued a very similar admonition in his 1943 poem, *The Secret of the Machines*:

> *But remember, please, the Law by which we live,*
> *We are not built to comprehend a lie,*
> *We can neither love nor pity nor forgive,*
> *If you make a slip in handling us you die!*

Of course, the sorcerer's apprentice motif, which Langdon Winner has called the "technics-out-of-control" theme (Hess, 1995, p. 371), is not restricted to AI and can be played out in several other domains of human endeavor (as can be readily intuited in the cases of genetic engineering, nuclear energy and politics).[8] We could even contend that a maneuver of the same ilk, albeit defanged from existential risk, is at play in the way in which social scientists will sometimes don the garbs of their counterparts in the *Naturwissenschaften*, a fact upon which Wiener himself heaps no little scorn:[9]

> The success of mathematical physics led the social scientist to be jealous of its power without quite understanding the intellectual attitudes that had contributed to this power. […] Just as primitive peoples adopt the Western modes of denationalized clothing and of parliamentarism out of a vague feeling that these magic rites and vestments will at once put them abreast of modern culture and technique, so the economists have developed the habit of dressing up their rather imprecise ideas in the language of the infinitesimal calculus. (1964, p. 90)

As in most instances of "cargo cult science"—the catchy label with which Richard Feynman (1985) has forever christened such cases in which only the outer trappings of a procedure are imitated while its essence is left utterly untapped—much to the bewilderment of our flummoxed apprentice and to the safety of our good green Earth, there is no bang for the true sorcerer to wrestle with, but merely an ineffectual whimper. What makes artificial intelligence terrifying in this respect, however, is the potential for power scaling that computers provide. Machines are not (or at the very least need not be) intrinsically evil and what they bring about will depend on how we humans play our cards:

> The computer is not a simple force for good […] but like all machines is just a lever, multiplying the power of whoever controls it. The computer will just as happily lend itself to the further enslavement, terrorizing, and deception of its users as it will to liberate, enlighten, and enrich them. (Halpern, 2008, ¶ 11)

As was pithily put by Eliezer Yudkowsky a researcher specializing in AI safety, value alignment and human rationality, into whose ideas we will delve in greater depth: "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" (2008a, p. 26).

Allen Newell, co-creator of two of the earliest AI programs, champions too the parallels between the power of intelligent computing and the magic that populates fairytales: "I see the computer as the enchanted technology. Better, it is the technology of enchantment. I mean that quite literally" (1992, p. 47). But he is far less pessimistic when commenting on Wiener's and others' gloomy forebodings, saying that the dangers have been exaggerated and that the rigidity of a machine's decision-making has been overstated. He focuses instead on the good that could come: "The aim of technology, when properly applied, is to build a land of Faerie […] computer technology offers the possibility of incorporating intelligent behavior in all the nooks and crannies of our world. With it we could build an enchanted land" (Newell, 1992, p. 47).

Meet the New Faith, Same as the Old Faith

And if the land of the faerie is at hand, as Newell posits, then what comes next? It turns out that the pace that takes us from fairies to genies and onwards to the gods is quite brisk, and we are suddenly confronting not merely the domain of fable and myth, but that of religion, too. Or, at the very least, its current secular and technophile incarnation. When science historian George Dyson (son of famed physicist Freeman Dyson[10]) was invited by Google to tour their campus on the sixtieth anniversary of John Von Neumann's death he felt a distinctively religious vibe floating around:

> My visit to Google? Despite the whimsical furniture and other toys, I felt I was entering a 14th-century cathedral—not in the 14th century but in the 12th century, while it was being built. Everyone was busy carving one stone here and another stone there, with some invisible architect getting everything to fit. The mood was playful, yet there was a palpable reverence in the air. "We are not scanning all those books to be read by people," explained one of my hosts after my talk. "We are scanning them to be read by an AI." (Dyson, 2005, ¶ 27)

The comparison between AI discourse and religious thought has been amply and explicitly addressed. In *The Religion of Technology*, historian David F. Noble argues that technology should not be seen as divorced from a religious heritage (as so many idolaters of a shallow scientism would have it) but rather deeply rooted in it and fulfilling the same primeval aspirations. He singled the case of AI as particularly salient:

> Artificial Intelligence advocates wax eloquent about the possibilities of machine-based immortality and resurrection, and their disciples, the architects of virtual reality and cyberspace, exult in their expectation of God-like omnipresence and disembodied perfection. […] All of these technological pioneers harbor deep-seated beliefs which are variations upon familiar religious themes. (Noble, 1999, p. 5)

Robert Geraci has explored these parallels at length, most notably in his 2012 book *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*, which opens with the strong claim that, other than fundamentalist Christian theologians, "popular science authors in robotics and artificial intelligence have become the most influential spokespeople for apocalyptic theology in the Western world" (Geraci, 2010, p. 8). In a similarly titled earlier paper he had already affirmed that:

> Apocalypticism thrives in modern robotics and AI. Though many practitioners operate on a daily basis without regard for the fantastic predictions of the Apocalyptic AI community, the advocates of

> Apocalyptic AI are powerful voices in their fields and, through their pop science books, wider culture. Apocalyptic AI has absorbed the categories of Jewish and Christian apocalyptic theologies and utilizes them for scientific and supposedly secular aims. (Geraci, 2008, p. 161)

AI is, however, not the first attempt to translate religious grand visions of the future into a goal that is within the grasp of science, technology and social reformation. Nanotechnology critic Lyle Burkhead (1997, ¶ 5), in discussing where extropianism[11] fits within the "memetic landscape" points out that despite having found rich soil in the current capitalist ecosystem, these ideas were already present in the ultimate ideals of Marxism:

> The basic Extropian vision, as I understand it, is that the whole world will be mechanized, the new transhuman species will emerge, and transhumankind will expand throughout space; and meanwhile the state will wither away.
>
> This is exactly comparable to the founding vision of the Soviet Union. Marx and the Bolsheviks weren't trying to establish a totalitarian state as an end in itself; the state was supposed to be a temporary thing that would eventually render itself unnecessary, and wither away. Meanwhile the whole world would be mechanized, and the New Communist Man would emerge. Space colonization wasn't part of the original vision, but it was implicit. […] The Bolsheviks were the first who had enough hubris to treat this as a practicable vision, something that could be made to actually happen. (Hubris has always been permitted; it's just that it has consequences.) Now, Extropians also want to make it actually, physically happen, but they want to do it within the capitalist economy. Instead of Karl Marx, their mentors are Robert Heinlein, Ayn Rand, Marvin Minsky, Vernor Vinge…[12]

This pursuit of AI as a means to "immanentize the eschaton"—to put it in Eric Voegelin's (1952) evocative phrasing, made immortal by William F. Buckley's vocal exhortation not to—is inextricably linked to transhumanism. Broadly speaking, the "transhumanism" label refers to a movement that seeks a departure from the limitations of being human and pursues extending our species' evolution through advanced technology in order to conquer death and enhance our all-too-feeble current organic minds and bodies.[13] This technological messianism appeals to our fantasy, making ample promises in a language poised somewhere between marketing and divination. In their view, humanity shall undergo a monumental transformation and leave behind our present state.

Among the current mentors pushing the transhumanist idea, probably none is as well-known and controversial as Ray Kurzweil. Considered the leading prophet (a term that both his followers and detractors would deem appropriate) of the advent of superhuman AI, he was appointed Director of

Engineering by Google in 2012 and along said company and the NASA Ames Research Center founded the Singularity University. With robotics pioneer Hans Moravec coming in a distant second place, Kurzweil is the main promoter for the advent of the Singularity, a concept which was originally coined by sci-fi author and computer scientist Vernor Vinge. (Which is just one among many examples showing how AI research and discourse feeds upon and responds to its treatment in fictional narratives, an idea which we'll explore at length in the next section.)

While the term 'Singularity' has been used with several distinct—albeit essentially related—meanings (Sandberg, 2010), it is basically understood as the point in history at which human intelligence, as it has existed ever since its evolutionary, biologic inception, will be radically surpassed by a new kind. David Chalmers (2010) defines it as:

> An intelligence explosion [with] enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more. It also has enormous potential dangers: an end to the human race, an arms race of warring machines, the power to destroy the planet. (Chalmers, 2010, p. 3)

And while there may be some other pathways that could lead to this outcome (such as genetic engineering, nanotechnology or mind uploading), AI is widely held as the most likely candidate, and is tacitly assumed as the cause when talking about the singularity (with the aforementioned routes occasionally touted as ancillary pathways to achieving it). Indeed, in popular discourse, public perceptions and mainstream fictional depictions, AI seems to be increasingly linked to the idea of an intelligence explosion.

Kurzweil pins the unavoidability of the advent of such a singularity on what he has called the "Law of Accelerating Returns". Inspired by Moore's Law, first identified by Intel's co-founder Gordon Moore, which (loosely restated) observes that computing power per dollar expended doubles roughly every 18 months, the LOAR claims that every single gain in computing technology on the way to superintelligence will compound, amounting to an exponential growth (Kurzweil, 2001).[14] To defend his position he explains that the expanse of time separating the crucial moments in this trajectory is diminishing exponentially. Thus, for instance, modern man and language appeared 1.400 generations ago. Writing goes back a measly 200, the printing press is from us but 20 generations removed and the computers have been with us for some two generations or so (Kurzweil, 1990).[15]

Many see in Kurzweil's predictions above all a desire for a salvation promised in robotic theology rather than something approaching scientific rigor. Ricardo Rosas (1992, p. 125, my translation) suspects that the latent reason for seeking to build artificial intelligences is precisely "the secret temptation of playing at being Gods" (though as we already mentioned, McCorduck would object). Far from denying any such claim, Kurzweil approvingly cites Ramez Naam when he defends that very drive. "'Playing God' is actually the highest expression of human nature. […] Without these urges to 'play God' the world as we know it wouldn't exist today" (Naam cited in Kurzweil, 2005, p. 299).

Kurzweil has certainly not been exempt from harsh criticism on the part of key characters in the field. In a 2017 interview, venerable forefather John McCarthy, who gave Artificial Intelligence its name (McCorduck, 1979), claimed that Kurzweil "has not provided any sufficient basis for his short term optimism." And in regards to the feasibility of Artificial Intelligence ever being achieved, McCarthy added that "maybe it will and maybe it won't, but if it does it won't be due to him" (Computer History Museum, 2017). Douglas Hofstadter, himself a maker of computer programs that model cognition and who has organized more than one panel on the topic of the Singularity calls the views of Kurzweil and Moravec "an intimate mixture of rubbish and good ideas, and it's very hard to disentangle the two, because these are smart people; they're not stupid" (Ross, 2007, ¶ 18). The vision of the Singularity, with its transcending of materiality and its arrival of a new world, espoused by Kurzweil and Moravec is often derisively termed "the rapture of the geeks" (DeBaets, 2015; Barrat, 2013).

> It's no surprise that the Singularity is often called the Rapture of the Geeks—as a movement it has the hallmarks of an apocalyptic religion, including rituals of purification, eschewing frail human bodies, anticipating eternal life, and an uncontested (somewhat) charismatic leader. (Barrat, 2013, p. 94)

Nevertheless, such a scornful dismissal misses the point and seems to be rather a handy way to avoid thinking about the issue and allaying our own uneasiness. In short, it's a stop sign for a serious analysis of the matter. Singularitarianism is not the awkward mongrel offspring of the faith of yore, as many would have it, but a full-fleshed and voracious descendent. After his rigorous analysis of the isomorphism of both discourses, Geraci concluded that "Apocalyptic AI is the legitimate heir to these religious promises, not a bastardized version of them" (2008, p. 158).

The utopia that Kurzweil is eagerly banking on, and that many technologists hope for chimes with the closing lines of this 1967 poem by Richard Brautigan:

I like to think
(it has to be!)
of a cybernetic ecology
where we are free of our labors
and joined back to nature,
returned to our mammal
brothers and sisters,
and all watched over
by machines of loving grace.

For all the devoted acolytes Newell's "land of Faerie" seems to have found, it must not be forgotten that the reverse side of the coin of a merciful omnipotent God is, as many a drowned character of Biblical lore could attest, a cruel omnipotent one or, almost as bad, an indifferent one. For as we'll see, when it comes to a superintelligence, the active thwarting of our goals and the oblivious ignoring of our plight are really not that different at all. Peter Thiel, the superstar entrepreneur we could (but probably don't) thank each time we make an online purchase (along Elon Musk) is the major donor of the Machine Intelligence Research Institute, one of the few institutions whose personnel is devoted to the preemptive forestalling of existential risk.[16] He has claimed that:

> Strong AI is like a cosmic lottery ticket: if we win, we get utopia; if we lose, Skynet substitutes us out of existence. (Thiel & Masters, 2014, p. 84)

The most serious and exhaustive analysis of what true risks a superintelligence entails we owe to Nick Bostrom, at the University of Oxford and founder of its Future of Humanity Institute. His thorough study on the topic, *Superintelligence: Paths, Dangers, Strategies*, deserves close consideration, for in a sober and rational way it addresses the real causes of concern regarding where the future of our technological innovations will lead us:

> [T]he prospect of superintelligence, and how we might best respond. This is quite possibly the most important and most daunting challenge humanity has ever faced. And—whether we succeed or fail—it is probably the last challenge we will ever face. (Bostrom, 2014, p. 7)

The sentiment is fully captured in the title of documentary filmmaker James Barrat's book *Our Final Invention: Artificial Intelligence and the End of the Human Era*, who after listening to an aging Arthur C. Clarke voicing his concerns that humanity would be superseded, set out to interview several of AI's

leading thinkers and main actors, in order to address the risk that smarter than human artificial intelligence would pose us a serious existential threat. "Before, I had been drunk with AI's potential. Now skepticism about the rosy future slunk into my mind and festered" (Barrat, 2013, p. 8). The phrase he chose as a title goes all the way back to 1966, when I.J. Good, a British mathematician who worked alongside Alan Turing to decipher German codes during the Second World War, wrote his seminal paper on the intelligence explosion:

> [T]he first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously. (Good, 1966, p. 33)

But just *how* seriously? A whole thesis could be written solely on the limits that should be imposed on drawing real world conclusions from the world of literature. And as should be expected, it turns out that there are good reasons both to support this role of fiction and to be wary of it.

Fiction as *Gedankenexperiment*

Artificial Intelligence research has been criticized in the past as being nothing but science fiction (Taube, 1961), but while such criticism is unduly harsh, undeniable feedback loops exist between both domains. In their exhaustive analysis of the visions of AI presented in the New York Times over the last 30 years, Fast & Horvitz (2017, p. 4) observe that while AI and science fiction have always been associated, that association pointedly increased at the start of the 90s. It is well and good to try to throw some clarity upon the murky waters of what may seem to be implied by our previous statement regarding feedback loops. Certainly, there is an important conceptual distinction to be made between the technical papers dealing with the rigorous detail of circuitry and programming, the pop science books chronicling the advent of thinking machines and the tales spun by scribblers of a speculative persuasion. However, the idea of AI has been relayed culturally to the general public, not through the domain of technical discussion, but by way of popular culture, be it in the form of movies, games books or TV shows. This zone of free access, more or less available to all, is where most people derive their notions of what AI is and is not.

The interrelatedness of AI and fiction can be understood in several different ways.[17] Fiction has an impact on the genesis, culture and future of artificial intelligence by virtue of being experienced (or lived-in) by its researchers; developments in AI affect the production of literature and other forms of

the fictional, and finally, narrative depictions of AI that are current in various media affect the attitudes of the public and the decisions made by policy-makers (Cave, Coughlan & Dihal, 2019; Fast & Horvitz, 2017). However, in practice, the boundaries between them are blurry.

As to the first proposal, much of what we have already explored of the past roots of AI can attest. Pamela McCorduck, when explaining what inspired her to capture the living story of the field, told by its very founders "before mortality claimed them" (McCorduck, 2004, p. xi) stated that she wanted her fellow humanists "to see a science whose genesis was in literary texts they cherish" (McCorduck, 1979, p. xix). Furthermore, it is almost impossible to conceive of a contemporary AI researcher who was not been reared on the fantasies of Isaac Asimov and Arthur C. Clarke.[18] The case of Clarke merits a little more detail, for not only has his HAL 9000 arguably become one of the most recognizable fictional AIs of all time, to the point that mention of his work is nay unavoidable in current discussions of AI, but also because he represents a perfect example of a writer working on the vanguard where science fiction meets science fact. He consulted with IBM in matters computer-related, although he has vehemently denied as groundless the rumors claiming that H A L was a one-step alphabetical transliteration of I B M (Clarke, 2000). Even more importantly, Marvin Minsky was a consultant for the movie *2001: A Space Odyssey*, and his theories and experimental results are explicitly mentioned in the novel that Clarke developed concurrently with the film's script (Minsky, 2007). In terms of inspiring ever-newer generations of wanderers of landscapes spatial and mental, what was the case for aspiring astronauts is still the case for AI researchers.

> How many young students have been "turned on" to science by reading science fiction? Most of the men who have walked on the Moon's surface trace their careers back to early readings in science fiction[19]. (Bova, 1974, p. 9)

The connection is so strong and evident that some researchers have even proposed a curriculum that purposefully employs science fiction as an entry point for the teaching of artificial intelligence and computer science to college students (Goldsmith & Mattei, 2014; Tambe, Balsamo & Bowring, 2008; Bates, 2011). Robert Geraci noted the strength of the link when he visited the Robotics Institute of Carnegie Mellon University[20], trying to better understand what had led Hans Moravec to write his pop-science books:

> Concerns about the military were relatively rare but interest in science fiction was commonplace. Although few researchers proposed that robotics or AI research might arise directly from science

> fiction or that there was a definite relationship between sci-fi and Apocalyptic AI, the genre came up in nearly every conversation I had (sometimes at my instigation but far more often not). The writers Isaac Asimov, Philip K. Dick, and Neal Stephenson and several TV shows and movies were all brought up by grad students, faculty, and researchers. Science fiction has a persistent presence in the lives of the RI faculty and students, so it takes little imagination to appreciate how it might affect the ideology of Apocalyptic AI. (Geraci, 2010, p. 41)

But driving them to the field in the first place is far from the only impact that narrative fiction has over AI researchers. As we have already hinted, the stuff of story land—like it or not—influences the thinking about AI that gets done, for much like other "semiotic resources" (Kress, 2010; Van Leeuwen, 2004), they act as anchoring points in idea-space. Douglas Hofstadter and Emmanuel Sander have likened a person engaging in the act of thinking while taking advantage of the vast conceptual storehouse offered up by the culture to a rock climber who follows the trail opened up by free-soloing pioneers:

> We who are alive today are the beneficiaries of countless thousands of conceptual pitons that have been driven into the metaphorical cliffs of highly abstruse situations. We can easily climb up steep slopes of abstraction that would have seemed impossible a few generations ago, for we have inherited a vast set of concepts that were created by ingenious forebears and that are easy to use. (Hofstadter & Sander, 2010, p. 131)

One archetypal such conceptual piton is the fable, which after being heard and internalized becomes an idealized abstraction readily available to be called upon for judging future situations and quickly deciding how to act:

> It becomes a label that jumps to mind when someone who has incorporated it in their memory runs into a situation that "matches" or "fits" the fable — not in a word-for-word fashion, obviously (fables are seldom memorized), but by an abstract alignment with its moral, or with its title, or just with a blurry memory of its basic plot. (Hofstadter & Sander, 2010, p. 111)

What's sauce for the goose is sauce for the gander and what's true for Mother Goose is also true for a robotic starship commander. Fables, fairytales, myths and science-fiction stories or novels function as a higher order language. If words aid us in crystallizing phenomena, carving up perceptible portions of the world and making it possible to communally transmit and share information about them, then art forms expand these powers of communication to even greater heights. Just like an emotion is shorthand designed by evolution for a complex string of survival-relevant thinking and decision-

making, works of art function as shorthand for culturally transmissible sequences of ideas and emotions. They summarize complex phenomena in an expressive way and by providing us with an indexical name, allow us to quickly refer to and confer upon them. A story becomes abstracted and its label suffices to evoke its structure, allowing us to even think in advance about things that haven't happened, but could.

> Pieces of fiction are simulations of selves in the social world. Fiction is the earliest kind of simulation, one that runs not on computers but on minds. One of the virtues of taking up this idea from cognitive science is that we can think that, just as if we were to learn to pilot an airplane we could benefit from spending time in a flight simulator, so if we were to seek to understand better our selves and others in the social world, we could benefit from spending time with the simulations of fiction in which we can enter many kinds of social worlds, and be affected by the characters we meet there. (Oatley, Mar & Djikic, in press, p. 4)

Science fiction in particular looks admirably well suited to the purpose of letting AI researchers run their mental simulations, for it provides them with a fertile and vivid playground for such hypotheticals as could inspire their theorizing:

> The science fiction writer is in the truest sense a professional fabricator of *gedankenexperimenten*, whether he is exploring the narrow consequences of a new scientific or technological development or whether he is considering the broader consequences of a social trend. (Scortia, 1974, p. 78)

Not only that, but it is also attuned to the background religious sensibilities that we have already noted, for, just like AI, "the sacralizations of space and technology of SF have reinvented 'religion' to fit the secular experiences of modern people" (Pels, 2013, p. 214). Both in AI as in sci-fi, Science with a capital S tried to fill in the gaping void left by the departure of God. "Science meets the specifications for a deity more than any other single thing in the current cultural cosmos", says science fiction author Theodore Sturgeon[21], given that it "presents all the attributes of an object of worship, and is accordingly respected, feared, sacrificed to, and invoked—that is to say, worshiped" (Sturgeon, 1974, p. 59).[22]

So far, so good, but what about the negative consequences of relying on fiction to inform our ideas of AI? Asimov's three laws of robotics are ubiquitous and nearly unavoidable but are they really what we should be currently paying attention to? They were first proposed in 1942, have we made no progress whatsoever since then in the programming of safety protocols for thinking machines?[23] And

does the fact that journalists writing on tech have seen *The Matrix* imply that it is a good idea for them to include comparisons to such movie scenarios in their every discussion of AI risk? Nick Bostrom is puzzled and vexed that when it comes to this particular topic, films and stories should always be discussed:

> There's a tendency to assimilate any complex new idea to a familiar cliché. And for some bizarre reason, many people feel it's important to talk about what happened in various science fiction novels and movies when the conversation turns to the future of machine intelligence. (Bostrom, 2015, p. 126)

Eliezer Yudkowsky (2007a) warns against what he has called the Logical Fallacy of Generalization from Fictional Evidence. The mere existence and box-office appeal of the *Terminator* movies should not, by any means, lead to said franchise being used as a starting point for most policy discussions of AI. This reliance on fictions can have a pernicious effect by making us unduly focus on too narrow a segment of probability space, biasing us to pay more attention to some scenarios than they actually merit, while downplaying the true risks of some other future outcomes that may be either more likely or more dangerous. The "seen" boogeyman could be far more benign than the one we fail to notice. A steel humanoid skeleton walking around with a machine gun is more cinematic than small nanoparticles that multiply by consuming all available matter lying nearby, but the damage that the latter could cause is unquantifiably greater than the former's. Forget anthropoid robot mercenaries, what's really terrifying are self-replicators with a warped utility function. For Yudkowsky (2011) and Bostrom (2003), one of the most egregious members of this class is the now infamous paperclip maximizer, an entity proposed by the latter which, just like Wiener warned, would blindly pursue its ill-stated goal of making as many paperclips as possible with complete disregard for the consequences of its relentless obsession, even if these included the total obliteration of all life in the universe.[24] If we don't succeed in properly instilling adequate values in the first self-improving superintelligence, Yudkowsky claims, "the result would not be a ghost-in-the-machine free to go its own way without our nagging, but a future light cone tiled with paperclips" (Yudkowsky, 2011, p. 14).

But the siren calls of fiction are too sweet and Yudkowsky himself[25], the Logical Fallacy of Generalization from Fictional Evidence notwithstanding, alludes to Greg Egan's (1997) novel *Diaspora* as the starting point for the thoughts that led him to formulate the idea of Coherent Extrapolated Volition, an attempt at setting in place meta-level guidelines for programming value-alignment into an AI in such a way that they could survive regardless of the exponential self-optimization that the AI underwent and remained in line with humanity's best interest, without, however, rigidly specifying in

advance what that best interest must be (Yudkowsky, 2004). Even more so, the paper in which CEV is outlined uses not one but two sci-fi novellas as examples of what end results should be avoided in such an attempt.

AI safety researcher Kaj Sotala when commenting on Yudkowsky's denouncing of the generalization from fictional evidence wondered whether alluding to *The Metamorphosis of Prime Intellect* (Williams, 1994), one of the two examples cited by Yudkowsky on his discussion of CEV, was warranted on the basis that it provided "a fictional example of an AI whose 'morality programming' breaks down when conditions shift to ones its designer had not thought about" (2007, ¶ 2). There's a strong case to be made that there is a difference between a fictional example which is purposefully chosen for a specific reason and one that is ready-made, just lying around and which we let come unbidden. This is much like with Orwell's clichés that force themselves upon our minds and therefore prevent our thinking. As he warns, the cost of "letting the ready-made phrases come crowding in" lies in that they "will construct your sentences for you — even think your thoughts for you, to a certain extent" (1946, ¶ 18).

Instinctive appeals to pre-digested scenarios would appear to be the problem, not the use of fiction per se; if there is an act of volition involved, then it is fair game to refer to fictional semiotic resources, it would seem. But even in that case, when the same starting point has been trodden over and over and over, it is difficult to reach new conclusions. And despite having acquiesced with the best of intentions to the cognitive expenditure of careful and judicious choice, we still submit ourselves to the risk posited by the cognitive bias of vividness. No matter the disjunctive probabilities of piling fact atop new shiny fact, the added details simply make us perceive engagingly-described scenarios as more plausible (Yudkowsky, 2008b). And this is particularly worrisome if we take into account that authors (*pace* Jules Verne), as entertaining and thought provoking as they may be, lack a consistent track record as accurate forecasters:

> There are basic incompatibilities between good story telling and accurate prophecy. A good story needs conflict and dramatic tension. […] The track record of SF writers as prophets, operating within these constraints, has not been impressive. The future, as has emerged, has rarely borne much resemblance to the near-future SF that preceded it. (Cramer, 1990, ¶5)

But while we must suppose professional researchers to be relatively protected in this respect, the same is not true of the public at large, which is a growing source of concern for policy makers political and

military. In a report on the ethical considerations of autonomous military robots prepared for the US Navy, Lin, Bekey & Abney (2008) identify public perceptions as one of the main market forces that are currently impacting the development of military robotics:

> From Asimov's science fiction novels to Hollywood movies such as Wall-E, Iron Man, Transformers, Blade Runner, Star Wars, Terminator, Robocop, 2001: A Space Odyssey, and I, Robot (to name only a few, from the iconic to recently released), robots have captured the global public's imagination for decades now. But in nearly every one of those works, the use of robots in society is in tension with ethics and even the survival of humankind. The public, then, is already sensitive to the risks posed by robots—whether or not those concerns are actually justified or plausible—to a degree unprecedented in science and technology. Now, technical advances in robotics are catching up to literary and theatrical accounts, so the seeds of worry that have long been planted in the public consciousness will grow into close scrutiny of the robotics industry with respect to those ethical issues, e.g., the book *Love and Sex with Robots* published late last year that reasonably anticipates human-robot relationships. (Lin, Bekey & Abney, 2008, p. 9)

They are rightly concerned about what direction the tides of the summer blockbusters may sway the willing audiences, for as cultural psychologist Jaan Valsiner has pointed out, "fictional characters have real consequences for humans living and dying on the battlefields – not just for the queries of readers of sophisticated novels" (2009, p. 101). Undeniably, works of fiction can have a sizable impact on the real world acting as cautionary tales and in that capacity, contributing to forestall some outcomes or subtly nudge us towards others:

> When you think about it, you realize these two works have influenced our world. Neither Brave New World nor 1984 will prevent our becoming a planet under Big Brother's thumb, but they make it a bit less likely. We've been sensitized to the possibility, to the way such a dystopia could evolve. (Herbert, 1974, p. 42)

We will offer one final example, due to the noteworthiness of its driving force, of a fictional scenario contingently impacting not only public perceptions of AI, but the attitudes and behaviors of the researchers themselves: the notion of Roko's Basilisk. Although purely speculative and up until this point nothing more than an imaginary entity, Roko's Basilisk is having an effect on part of the community of friendly AI researchers, particularly the rationalists working on existential risk, to the extent that it has been deemed a dangerous idea and the mere mention of it has been strongly discouraged. What could make a purely fictional creature so terrifying and so worthy of these

cautionary measures? Roko's Basilisk is a hypothetical future artificial superintelligence, that, if it came into existence, would retroactively institute, through coercion, the set of policies that would have hastened its coming into existence. More concretely put, it is presumed to be so powerful as to be able to torture all those who knew of the possibility of its eventual existence, but did no invest a significant amount of their efforts and resources to actualizing its potential. Not even death would be a safeguard against this nightmarish scenario, as the Basilisk is presumed to be so advanced as to be able to create perfect simulations of the transgressing researchers which it would eternally punish. Far-fetched? Most certainly, and yet there's no denying that this egregore, this collective mental entity, has a certain psychological pull, and that many who have learned of the concept dearly wish they'd never heard of it.

Denouement

I have attempted to highlight the interrelatedness of literary fiction, myth and religion with the theorizing and dissemination of AI ideas by a significant percentage, if not a majority, of its practitioners, trying to portray through picturesque examples the underlying connection to ancestral human motivations that drive researchers in their pursuit, but that, more generally, fascinate the public and humanity at large. There are good reasons for exploring these points of narrative entanglement where AI meets the wider culture and draws from it its vital sap, other than the sheer fun and delight of reading about such things. Latour (1987) foundationally opened our eyes to the importance of studying scientist in the true expanse of their ecosystem, paying attention not only to their published output but to the culture they were a part of, for it brings out a fuller picture which can enrich our understanding of a field. More recently, Arthur Melzer (2007) has made a very well grounded case for how teachings in mainstream science are not always transmitted overtly, but oftentimes through esoteric means. Some fables functioned in the past like veritable samizdats, disguising knowledge and moving it past censors, and in a similar fashion, we could argue that what gets passed on today about AI is not solely contained in handbooks and papers, but in novels and films as well. These stories feed the argumentational promise (Barutta, Cornejo & Ibáñez, 2011) of Artificial Intelligence, that is, a tacit commitment driving researchers in their quest to expand the discipline.[26] In this light, it does become important to pay attention to the lore, ancient and modern, surrounding AI research.

However, such parallelisms have been outlined as a way to render even more visible the aesthetic attractiveness of the topic so as to draw attention to it on the part of newer audiences, and in no way

should they be seen to invalidate the very real concerns of those who are leading the discussion of existential risk associated to AI as childish speculation that results from the consumption of too many a science fiction novel (even if some of the most extreme beliefs in that sphere, such as Roko's Basilisk could seem outlandish at first), but rather as a call to engage ourselves with that discussion and raise awareness as to the potential destination of this technology. No matter how steeped the language of Artificial Intelligence may be in the religious and mythical traditions or in the accumulated wealth of the science fiction canon and how much vividness it may derive from them, it would be a grievous blunder to irresponsibly disregard the feasibility of higher than human level intelligence eventually being attained by machines.

> So if there is even a small chance that there will be a singularity, we would do well to think about what forms it might take and whether there is anything we can do to influence the outcomes in a positive direction. (Chalmers, 2010, p. 3)

Unfortunately, what should be addressed in sober and technically accurate terms will more often than not reach a wider audience through sensationalist and sloppy reporting.[27] This is extremely problematic since, given enough of these "scary reports", much like the once trusting co-villagers of the boy who cried wolf, people will begin developing a resistance to serious calls for concern that are actually grounded in what is truly going on. And just as there are narrow-minded reasons to exaggerate AI's current risks and achievements there are and have been wider social reasons, military and economical, to downplay them. The widely disseminated idea that computers were nothing but "fast morons", strictly incapable of doing anything but what they were ordered, was a deliberate marketing move on the part of computer vendors in order to ease buyers into bringing the then-novelty device into their homes (McCorduck, 1979, p. 202).

Academics are a part—or should aspire to be—of a stigmergic network that slowly accrues value in its insights. Therefore, even if it may seem liable to invite superficial groupthink to claim that ideas that have gained more traction should be prioritized, there is a point to be made for the attention owed to the laborious unearthing of choice paragraphs in the works of primary sources. If this were not the case, and leaving aside the importance of visiting the classics personally rather than relying on secondary commentators, all of the endeavors of literary and academic critique and analysis would be vain. Mustering what powers and platforms of communication one can summon to amplify a distress signal is a warranted ethical move.

As so many of us have had to learn from baseball catcher-cum-philosopher Yogi Berra's attributed wisdom, predictions are especially hard when they involve the future. Let us, before departing, pay one final visit to Newell's fairyland and ponder his admonition in the face of uncertainty, which rings today truer than ever:

> The experts notwithstanding, fairy stories are for all of us. Indeed, this is true, if for no other reason than that today, we are all of us children with respect to the future. We do not know what is coming. It is as new to us and as incomprehensible as adult life is to a child. (Newell, 1992, p. 46)

References

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, March, 1942.

Barrat, J. (2013). *Our Final Invention: Artificial Intelligence and the End of the Human Era*. Chicago: Thomas Dunne Books.

Barutta, J., Cornejo, C. & Ibáñez, A. (2011). Theories and Theorizers: A Contextual Approach to Theories of Cognition. *Integrative Psychological and Behavioral Science*, 45(2), 223-246. ISSN: 1932-4502.

Bates, R.A. (2011, June), *AI & SciFi: Teaching Writing, history, Technology, Literature, and Ethics*. Paper presented at 2011 ASEE Annual Conference & Exposition, Vancouver, BC. https://peer.asee.org/17433

Bova, B. (1974). The Role of Science Fiction. In R. Bretnor (Ed.), *Science Fiction Today and Tomorrow*. NY: Harper & Row.

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. In I. Smit (Ed.) *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, pp. 12-17.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, United Kingdom: Oxford University Press.

Bostrom, N. (2015). It's Still Early Days. In J. Brockman (Ed.), *What to Think About Machines That Think*. NY: HarperCollins. pp. 126-127

Bradley, T. (2017, July 31). Facebook AI Creates Its Own Language In Creepy Preview Of Our Potential Future. *Forbes*. Retrieved from: https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future

Brautigan, R. (1967). *All Watched Over by Machines of Loving Grace*. San Francisco, CA: The Communication Company.

Burkhead, L. (1997). Extropianism in the Memetic Ecosystem. *Extropians Message Board*.

Butler, S. (1863). *Darwin Among the Machines*. Christchurch Press, June 13. Retrieved from: http://www.nzetc.org/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html.

Cave, S. & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2), 74.

Cave, S., Coughlan, K., & Dihal, K. (2019, January). 'Scary robots': examining public responses to AI. In *Proc. AIES* http://www. aies-conference. com/wp-content/papers/main/AIES-19_paper_200. pdf.

Chalmers, D. J. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*. 17:7-65. Retrieved from: http://consc.net/papers/singularity.pdf

Clark, S. R. L. (1995). Tools, Machines, and Marvels. In R. Fellows (Ed.), *Philosophy and Technology*. Cambridge: Cambridge University Press.

Clarke, A. C. (2000). *2001: A Space Odyssey*. NY: ROC.

Collins, T. (2017, July 31). Facebook shuts down controversial chatbot experiment after AIs develop their own language to talk to each other. *Daily Mail*. Retrieved from: https://www.dailymail.co.uk/sciencetech/article-4747914/Facebook-shuts-chatbots-make-language.html

Computer History Museum. (2017, 9, 11). Oral History of John McCarthy [Video file]. Retrieved from: www.youtube.com/watch?v=KuU82i3hi8c

Comrada, N. (1995). Golem and Robot: A Search for Connections. *Journal of the Fantastic in the Arts* 7(2/3) pp. 244-254.

Cornejo, C., & Musa, R. (2017). The physiognomic unity of sign, word, and gesture. *Behavioral and Brain Sciences, 40*, E51. doi:10.1017/S0140525X15002861

Cramer, J. G. (1990). Technology Fiction (Part I). *Foresight Update 8,* March 15. Retrieved from: http://www.islandone.org/Foresight/Updates/Update08/Update08.2.html

DeBaets, A. M. (2015). Rapture of the Geeks: Singularitarianism, Feminism, and the Yearning for Transcendence. In C. Mercer & T. J. Trothen (Eds.) *Religion and Transhumanism*. CA: Praeger. pp. 181-197.

Deutsch, D. (2011). *The Beginning of Infinity*. London: Allen Lane.

Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, Mass: MIT Press.

Dryden, J. (1913). *The Poems of John Dryden*, ed. by John Sargeaunt. London, New York: Oxford University Press. Retrieved from: https://www.bartleby.com/204/199.html

Dyson, F. (2015). I Could Be Wrong. In J. Brockman (Ed.), *What to Think About Machines That Think*. NY: HarperCollins. pp. 126-127

Dyson, G. (2005). Turing's Cathedral. *Edge*. Retrieved from: www.edge.org/conversation/george_dyson-turings-cathedral

Egan, G. (1997). *Diaspora*. London: Orion.

Eliot, G. (1997). *The Mill on the Floss*. Hertfortshire: Wordsworth Editions.

Eliot, T. S. (2011). [Letter written December 31, 1914 to Conrad Aiken]. In *The Letters of T.S. Eliot: Volume 1*. London: Faber and Faber.

Fast, E., & Horvitz, E. (2017, February). Long-term trends in the public perception of artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*. Preprint at https://arxiv.org/abs/1609.04904 (2016).

Fellows, R. (1995). Welcome to Wales: Searle on the Computational Theory of Mind. In R. Fellows (Ed.), *Philosophy and Technology*. Cambridge: Cambridge University Press.

Feynman, R. (1985). *Surely You're Joking, Mr. Feynman! Adventures of a Curious Character*. Bantam Books.

Foerst, A. (2004). *God in the Machine: What Robots Teach Us About Humanity and God*. NY: Dutton.

Geraci, R.M. (2008). Apocalyptic AI: Religion and the promise of artificial intelligence. *Journal of the American Academy of Religion*, 76(1), 138-166.

Geraci, R.M. (2010). *Apocalyptic AI: Visions of heaven in robotics, artificial intelligence, and virtual reality.* Oxford: Oxford University Press.

Goldsmith, J. & Mattei, N. (2014). Fiction as an introduction to computer science research. *ACM Transactions on Computing Education* (TOCE), 14(1), 4.

Good, I.J. (1966). Speculations concerning the first ultraintelligent machine. *Advances in Computers* (Vol 6) Chicago: Elsevier, pp. 31–88

Griffin, A. (2017, July 31). Facebook's Artificial Intelligence Robots Shut Down After They Start Talking to Each Other in Their Own Language. The Independent. Retrieved from: https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html

Halpern, M. (2008). The Trojan Laptop. *Vocabula Review* Vol. 10 Issue 1. Retrieved from: http://www.rules-of-the-game.com/com007-trojanlaptop.htm

Harrison, H. & Minsky, M. (1992). *The Turing Option*. NY: Warner.

Herbert, F. (1965). *Dune*. Philadelphia: Chilton Books.

Herbert, F. (1974). Science Fiction and a World in Crisis. In R. Bretnor (Ed.), *Science Fiction Today and Tomorrow*. NY: Harper & Row.

Hess, D. J. (1995). On Low-tech Cyborgs. In C. Hables Gray, H. Figueroa-Sarriera & and S. Mentor (Eds.), *The Cyborg Handbook*. New York: Routledge. pp. 371-78.

Horgan, J. (2016, March 1). AI Visionary Eliezer Yudkowsky on the Singularity, Bayesian Brains and Closet Goblins. *Scientific American*. Retrieved from: https://blogs.scientificamerican.com/cross-check/ai-visionary-eliezer-yudkowsky-on-the-singularity-bayesian-brains-and-closet-goblins/

Hurtado, E. (2017). Consequences of Theoretically Modeling the Mind as a Computer. (Doctoral dissertation). Pontificia Universidad Católica de Chile.

Ingold, T. (2007) *Lines: A brief history*. Oxon, UK: Routledge.

James, W. (1879). Are We Automata? *Mind, 4*, 1-22. Retrieved from: http://psychclassics.yorku.ca/James/automata.htm

Johnson, G. (1998). Science and Religion: Bridging the Great Divide. *New York Times*. Retrieved from http://www.nytimes.com/1998/06/30/science/essay-science-and-religion-bridging-the-great-divide.html

Kline, R. (2011). Cybernetics, automata studies, and the Dartmouth conference on artificial intelligence. *IEEE Annals of the History of Computing*, 33(4), 5-16.

Kress, G. (2010). *Multimodality*. London: Routledge.

Kurzweil, R. (1990). *The Age of Intelligent Machines*. Cambridge, Mass: MIT Press.

Kurzweil, R. (2001). The Law of Accelerating Returns. *KurzweilAI.net*. Retrieved from: https://www.kurzweilai.net/the-law-of-accelerating-returns

Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. NY: Penguin.

LaGrandeur, K. (2003). *Magical Code and Coded Magic: The Persistence of Occult Ideas in Modern Gaming and Computing*. Paper presented at the Conference of the Society for Literature, Science and the Arts. Retrieved from: http://ieet.org/index.php/IEET/more/lagrandeur20131026

LaLancette, HP (2007). *The Law of Accelerating Toilet Brushes*. Retrieved from: http://blog.infeasible.org/2007/05/29/the-law-of-accelerating-toilet-brushes.aspx

Lancaster, B. L. (1997). The Golem as a Transpersonal Image: A Marker of Cultural Change. Transpersonal Psychology Review 1, no. 3. pp. 5 – 11

Lancaster, B. L. (2007). La esencia de la Kábala: La enseñanza interior del Judaísmo. EDAF.

Latour, B. (1987). *Science in Action*. Cambridge, MA: Harvard University Press.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D. & Batra, D. (2017). Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint* arXiv:1706.05125.

Lin, P., Bekey. G. & Abney, K. (2008). *Autonomous Military Robotics: Risk, Ethics, and Design*. San Luis Obispo, CA: California Polytechnic State University.

Markoff, J. (1992, April 12). Technology; A Celebration of Isaac Asimov. *The New York Times*.

McCarthy, J. (2014). The Robot and the Baby. In D. H. Wilson & J. J. Adams (Eds.) *Robot Uprisings*. London: Simon & Schuster, pp. 343-362. Retrieved from: http://www-formal.stanford.edu/jmc/robotandbaby/robotandbaby.html

McCorduck, P. (1979). *Machines Who Think*. San Francisco: W. H. Freeman and Company.

McCorduck, P. (2004). Foreword to *Machines Who Think*. Natick, Massachussets: A K Peters.

McDermott, D. (1981). Artificial intelligence meets natural stupidity. In J. Haugeland (Ed.) *Mind Design*. Cambridge: Mass, MIT Press, pp. 143-60.

Melzer, A. (2007). On the Pedagogical Motive for Esoteric Writing. *The Journal of Politics*, Vol. 69, No. 4, November 2007, pp. 1015–1031.

Minsky, M. (2007, November 14). *Scientist on the Set: An Interview with Marvin Minsky*. Retrieved from: https://web.archive.org/web/20071113031417/http://mitpress.mit.edu/e-books/Hal/chap2/two3.html

Musa, R., Olivares, H. & Cornejo, C. (2015). Aesthetic aspects of the use of qualitative methods in psychological research. In G. Marsico, R. A. Ruggieri, & S. Salvatore (Eds.), *Reflexivity and psychology* (pp. 87-116). Charlotte, NC: Information Age.

Newell, A. (1992). Fairy Tales. *AI Magazine*, Volume 13, Number 4, pp. 46–48.

Noble, D. F. (1999). *The Religion of Technology: The Divinity of Man and the Spirit of Invention*. NY: Penguin.

Oatley, K., Mar, R. A., & Djikic, M. (in press). The psychology of fiction: Present and future. En I. Jaén & J. Simon (Eds.), *The Cognition of Literature*. New Haven, CT: Yale University Press.

Omohundro, S. M. (2008). The Basic AI Drives. In P. Wang, B. Goertzel & S. Franklin (Eds.) *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS

Orwell, G. (1946). Politics and the English Language. *Horizon,* volume 13, issue 76, pp. 252–265.

Pels, P. (2013). Amazing Stories: How Science Fiction Sacralizes the Secular. In J. Stolow (Ed.), *Deus in Machina: Religion, Technology, and the Things in Between*. NY: Fordham University Press.

Polanyi, M. (1983). *The Tacit Dimension*. Gloucester: Peter Smith Publisher, Inc.

Rosas, R. (1992). ¿Comerán los androides el fruto prohibido? Reflexiones acerca del Test de Turing. *Apuntes de Ingeniería* 45 (1992) pp. 111-129.

Ross, G. (2007). An interview with Douglas R. Hofstadter. *American Scientist*

Sandberg, A. (2010). An overview of models of technological singularity. In *Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March* (Vol. 8).

Scortia, T. N. (1974). Science Fiction as the Imaginary Experiment. In R. Bretnor (Ed.), *Science Fiction Today and Tomorrow*. NY: Harper & Row.

Shelley, M. (1818). *Frankenstein; or, the Modern Prometheus*. London: M. K. Joseph.

Sotala, K. (2007, June 17). *The Logical Fallacy of Generalization from Fictional Evidence* [Blog comment]. Retrieved from: https://www.lesswrong.com/posts/rHBdcHGLJ7KvLJQPk/the-logical-fallacy-of-generalization-from-fictional#gchLRgHocaajGkEy2

Sturgeon, T. (1974). Science Fiction, Morals, and Religion. In R. Bretnor (Ed.), *Science Fiction Today and Tomorrow*. NY: Harper & Row.

Tambe, M., Balsamo, A. & Bowring, E. (2008). Using science fiction in teaching artificial intelligence. *AAAI Spring Symposium*, 86–91.

Taube, M. (1961). *Computers and Common Sense: The Myth of Thinking Machines*. NY: Columbia University Press.

Thiel, P. & Masters, B. (2014). *Zero to One: Notes on startups, or how to build the future.* New York: Crown Business.

Valsiner, J. (2009). Between fiction and reality: Transforming the semiotic object. *Sign Systems Studies* 37 (1/2). pp. 99–113

Van Leeuwen, T. (2004). *Introducing Social Semiotics: An Introductory Textbook*. London: Routledge.

Vico, G. (1948). *The New Science of Giambattista Vico*. (Translated by Thomas Goddard Bergin & Max Harold Fisch). Ithaca: Cornell University Press.

Voegelin, E. (1952). *The New Science of Politics*. Chicago: University of Chicago Press.

Whelan, D. (2015, March 2). The Harry Potter Fan Fiction Author Who Wants to Make Everyone a Little More Rational. *VICE*. Retrieved from: https://www.vice.com/en_us/article/gq84xy/theres-something-weird-happening-in-the-world-of-harry-potter-168

Wiener, N. (1964) *God & Golem, Inc.: A Comment on Certain Points Where Cybernetics Impinges on Religion.* Cambridge, MA: MIT Press.

Williams, R. (1994). *The Metamorphosis of Prime Intellect*. Retrieved from: http://localroger.com/prime-intellect/mopiidx.html

Yudkowsky, E. (2004). *Coherent Extrapolated Volition*. Berkeley, CA: Machine Intelligence Research Institute. Retrieved from: https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2007a, October 17). *The Logical Fallacy of Generalization from Fictional Evidence*. [Blog post]. Retrieved from: www.lesswrong.com/posts/rHBdcHGLJ7KvLJQPk/the-logical-fallacy-of-generalization-from-fictional

Yudkowsky, E. (2007b, November 2). *An Alien God*. [Blog post]. Retrieved from: www.lesswrong.com/posts/pLRogvJLPPg6Mrvg4/an-alien-god

Yudkowsky, E. (2008a). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom & M. M. Ćirković (Eds.) *Global Catastrophic Risks*. New York: Oxford University Press, pp. 308–345

Yudkowsky, E. (2008b). Cognitive Biases Potentially Affecting Judgment of Global Risks. In N. Bostrom & M. M. Ćirković (Eds.) *Global Catastrophic Risks*. New York: Oxford University Press, pp. 91–119

Yudkowsky, E. (2011). Complex Value Systems are Required to Realize Valuable Futures. In J. Schmidhuber, K. R. Thórisson & M. Looks (Eds.) *Artificial General Intelligence: 4th International*

Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings, pp. 388–393. Retrieved from: https://intelligence.org/files/ComplexValues.pdf

Yudkowsky, E. (2015). *Harry Potter and the Methods of Rationality*. Retrieved from: http://www.hpmor.com

Computing Machinery and the Benefit of the Doubt

Roberto Musa G.

"If we could find a potato that is Socrates' peer in performance,
then we should have, for all practical purposes, Socrates,
and the potato would not be a potato after all but Socrates.
In science things are what they do."

Edwin G. Boring, 1946, p. 174

Abstract

Turing's *Computing Machinery and Intelligence* is one of the sacred texts of AI and the imitation game therein described has been fertile enough to nourish generations of researchers throughout a debate spanning seven decades. Despite its enormous and undeniable influence, we believe the Turing Test runs the risk of being contemporarily considered mainly a quaint and lackluster footnote in the story of Artificial Intelligence. Drawing partially on the historical and biographical context surrounding its writing, this essay seeks to highlight how much we stand to gain, in terms of both scientific integrity and personal empathy, by receiving Turing's message of truth-seeking open-mindedness as laid out in his groundbreaking paper. We argue against some of the conventional criticisms that are routinely leveled against the Test (such a Searle's), and pay special attention to one of the less talked about aspects of Turing's paper: a hard to gauge and historian-puzzling appeal to extra-sensory perception, for which we offer an explanation and which we use as an illustration of Turing's commitment to give the benefit of the doubt even in those cases in which it was inconvenient to do so. We argue that by imbuing our outlook with Turing's ethos we protect ourselves from the danger of too eagerly trusting ready-made labels rather than the data brought in by our own senses, something that puts us at risk of eventually disregarding truly intelligent beings as mere unthinking mechanical contrivances, due to irrelevant factors or official decrees.

Keywords: Turing Test, Artificial Intelligence, empathy, solipsism, Chinese room

At the very midpoint of the twentieth century a venerable British journal published a commentary dealing with the subject of the existence of God, which was aptly titled *The Existence of God*. In it, Thomas McPherson argued against an attempt to render Christian doctrine in a way verifiable by human experience. In particular, he strongly objected that the proposition "God exists" be operationally translated as "Some men and women have had, and all may have, experiences called 'meeting God.'"

This piece of theological dispute would be otherwise unremarkable but for the fact that some hundred pages before it in the same issue ran an article by a young mathematician which made an appeal to take a similarly difficult to tackle, metaphysically-sized question and "replace [it] by another, which is closely related to it and is expressed in relatively unambiguous words."

The journal was *Mind,* and while Alan Turing's treatment of the potential existence of thought in machines does not quite rise (yet) to the level of controversy that has surrounded the existence of God, both his reframing of the question "Can machines think?[28]" and the working answer he gave in that seminal paper have provided generations of researchers, commentators and casual onlookers with something to embrace, contest, or merely ponder.[29]

Turing's Paper: Stage-setting and Excursus

So much has been written hence that in commenting on Turing's *Computing Machinery and Intelligence* or its subsequent reception it is as daunting to stride for originality as for exhaustivity.[30] It should not surprise us to find it described as "the most reprinted, cited, quoted, misquoted, paraphrased, alluded to, and generally referenced philosophical paper ever published" (Halpern, 2006, p. 42). What I seek to show in this piece is that, despite how familiar, almost banal, the Turing Test may seem to us now, on the one hand, we run the risk of misinterpreting it while playing it out in our heads, and on the other and more importantly, we may fail to see that far from serving only as a criterion to discern true machine intelligence, Turing has given us a precious roadmap with which to navigate social interactions with personal integrity, empathy and scientific responsibility.

The precise parentage of the computer is hotly debated (see, for instance, Burks 2003) but Turing's name is as good as they come when looking for a mythical *urvater* for the machines that we have so inextricably come to depend upon (the Cronus to Turing's Zeus, Charles Babbage can then occupy a

grandfatherly role in this hierarchy of invention). What is far clearer, however, is that Turing was among the first to carry out serious research to support the position that machines could eventually be able to think[31]. While the 1950 paper (henceforth *CMI*) contains Turing's best known and most thoroughly elaborated criterion[32] for ascertaining whether one of them does, there is evidence that he had been considering the issue since 1941 at the very least (Copeland, 2004, p. 353).

Turing insists on the importance of being precise when asking questions of this caliber, and given that "can something think" seems nebulous to him, he proposes to convert it into what he terms the "imitation game". Distilled to its essentials, Turing's argument seems to boil down to the belief that a machine able to converse with a qualified human observer in such a way as to be indistinguishable from a human being should count as a thinking entity, for all intents and purposes. Daniel Dennett (2004) has suggested that Turing may have been inspired by Descartes' positing, in his *Discourse*, intelligent conversation as the hallmark distinguishing men from machines and Abramson (2011) has offered evidence for this view from Turing's archive.

Academia and the entertainment industry have, each in their own way, done their best to render the test accesible to all, but in delving into the winding and branching paths that have sprouted from Turing's article over the decades, and the hermeneutic polemics that each new commentator adds to the bubbling mix, we are likely to lose our way and become disoriented. Therefore we will provide at the outset, as useful guiding signposts, two brief but authoritative descriptions thereof:

> [The test] posits putting a computer and a human in separate rooms and connecting them by teletype to an external interrogator, who is free to ask any imaginable questions of either entity. The computer aims to fool the interrogator into believing it is the human; the human must convince the interrogator that he/she is the human. If the interrogator cannot determine which is the real human, the computer will be judged to be intelligent. (French, 2012, p. 164)

> The imitation game involves three participants: a computer, a human interrogator, and a human 'foil'. The interrogator attempts to determine, by asking questions of the other two participants, which of them is the computer. All communication is via keyboard and screen, or an equivalent arrangement (Turing suggested a teleprinter link). The interrogator may ask questions as penetrating and wide-ranging as he or she likes, and the computer is permitted to do everything possible to force a wrong identification. (So the computer might answer 'No' in response to 'Are you a computer?' and might follow a request to multiply one large number by another with a long pause and a plausibly incorrect answer.) The foil must help the interrogator to make a correct identification. (Copeland, 2004, p. 434)[33]

Fully suspecting the opposition his views would raise, Turing preemptively addressed, in style, nine possible objections to his proposal. It is his reply to the fourth of these, the Argument from Consciousness, that shall constitute the meat and potatoes of this essay for, as I will argue, therein lies a seed of the greatest importance in our dealing with one another, as human beings. But before doing so I'll touch upon two others, the first and last, as a way to hopefully whet the appetite for the original in those readers who may be personally unacquainted with it. For *CMI* is deservedly a classic and there is no way to do justice to the full scope of a classic in summaries or commentaries; their very nature resists condensation. When approached, the outer shell, that thin veneer floating around endemically in our culture's lingo falls apart to reveal an inner world of unbearable richness. As Italo Calvino stated, classics will always be new and fresh because they will always surprise: "the more we think we know them through hearsay, the more original, unexpected, and innovative we find them when we actually read them" (2000, p. 13). Or, to translate, *à la* Shannon, into information theory terms, classics are those works least amenable to lossless compression.

The first reply is a tongue-in-cheek riposte to the so-called Theological Objection, namely, that machines cannot think because thought is the prerogative of souls and these are to be conferred exclusively by the Creator upon humans. Turing parries the protest of his would-be critics squarely, on their own terms:

> It appears to me that the argument quoted above implies a serious restriction of the omnipotence of the Almighty. […] should we not believe that He has freedom to confer a soul on an elephant if He sees fit? […] 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. (Turing, 1950, p. 443)[34]

The ninth objection in his list has baffled a sizable proportion of the readership: the Argument from Extra-Sensory Perception. (I will delve on this issue at greater length that would perhaps seem warranted, but I believe this justified for reasons twofold, in that by so doing I propose and support a likely explanation for this hitherto most puzzling passage and lay some groundwork that will at the end tie in with the main theme of my piece.) As we have seen, the key for the implementation of Turing's proposal was to have the human and the machine sealed apart from the evaluator in closed rooms, so that they could communicate exclusively through printed written outputs. But what if either the evaluator or the human subject were to be endowed with psychic powers, and were to use them

to detect or transmit—by means other than those intended by the test—which answerer is human and which the computer?

From our contemporary perspective, that Turing would even deign consider such a possibility is perplexing. Especially since it doesn't connect with anything else in the paper and is hardly mentioned elsewhere in Turing's writings. Yet he tackles it head on, and offers a curious "solution". He asserts (1950, p. 454) that "to put the competitors into a 'telepathy-proof room' would satisfy all requirements" (whatever that may be and however it may be constructed).

Now, why would Turing mention such a thing? It seems entirely out of place. Even Andrew Hodges, the most important of Turing's biographers is at a loss to explain it, calling the following extract in the ESP section of *CMI* "the strangest passage in all Turing's writing" (1997, p. 48):

> These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately the statistical evidence, at least for telepathy, is overwhelming. It is very difficult to rearrange one's ideas so as to fit these new facts in. (Turing, 1950, p. 453)

Hodges then comments that "it is not clear how serious the statements are. The exclamation mark suggests irony, the 'overwhelming' evidence sounds literal" (1997, p. 49). In another place, Hodges keeps pondering whether Turing meant what he wrote, but adds a new angle by saying that his willingness to give ESP the benefit of the doubt was due to his placing the weight of data before his preconceptions:

> Readers might well have wondered whether he really believed the evidence to be 'overwhelming', or whether this was a rather arch joke. In fact he was certainly impressed at the time by J. B. Rhine's claims to have experimental proof of extra-sensory perception. It might have reflected his interest in dreams and prophecies and coincidences, but certainly was a case where for him, open-mindedness had to come before anything else; what *was so* had to come before what it was convenient to think. (Hodges, 2000, p. 416)

In *Gödel, Escher, Bach: An Eternal Golden Braid*, Douglas Hofstadter argues in a similar vein that: "Turing was reluctant to accept the idea that ESP is real, but did so nonetheless, being compelled by his outstanding scientific integrity to accept the consequences of what he viewed as powerful statistical evidence in favor of ESP" (1979/2000, p. 599).

My encounter with this oddest of bits in *CMI* left me unsatisfied at the explanatory dearth to account for it, until years later I hit upon a likely explanation. When in 2014 a copy of the original October 1950 issue of *Mind* went up for auction, I, like surely so many others, stood longingly looking in awe at it from the financially judicious distance of my computer screen, when something caught my eye. The cover listed three editors for the publication, and alongside the two well-known names of Gilbert Ryle and Frederic Bartlett I spotted one C.D. Broad, totally unknown to me. This prompted a query through which I learned that besides being a co-editor of the magazine when Turing's *CMI* was published, Charlie Dunbar Broad was a vocal supporter of ESP and had even been, in 1935, President of the Society for Psychical Research. On a philosophical autobiography of sorts Broad claimed that:

> A great deal of so-called scepticism is simply a particular kind of dogmatism which leads men to reject all alleged facts which do not come within the sphere of recognized science. Mine is certainly not of that type. I have always been interested in the phenomena dealt with by Psychical Research, and the attitude of orthodox scientists towards them has always seemed to me ridiculous. This view has been strengthened by subsequent intercourse with the skeletons which inductive logic conceals in its cupboards. Thus my scepticism makes me far less ready to reject the abnormal than are most educated men of our time. A man must know a great deal more about the secrets of nature than I do to reject any alleged fact without investigation, however wild it may seem. (Broad, 1924, §3)

Furthermore, he was one of the few publicly homosexual intellectuals of his time[35]. If we consider how Turing's life ended, we can see why this could hint to a possible affinity between both thinkers. Naturally, I also wondered whether this could be one of those cases, familiar to anyone who has published an academic paper or attempted to do so, in which Reviewer 3 kindly suggests adding a last-minute mention of his unrelated but favorite hobby-horse to the submitted manuscript. Or maybe Turing had by his own initiative added the peculiar passage in deference to Broad. I thought this well-grounded speculation worthy of being further investigated by those with access to the archives of either man (to whose ranks, alas, I don't belong).

Then, in 2017, David Leavitt came out with a masterful attempt to thoroughly explain Turing's mention of ESP and he noted Broad as a likely suspect for Turing's interest in the paranormal:

> The most likely scenario is that Turing became acquainted with Soal's research when he was in Cambridge in 1947–48, possibly through the agency of C. D. Broad. A Fellow of Trinity College, Broad was an active member of the SPR, a friend of Bertrand Russell and G. H. Hardy and, like Turing, unapologetically gay. In 1945 he had published a strong endorsement of Soal's work in the journal

> *Philosophy*, describing it as providing 'evidence which is statistically overwhelming for the occurrence not only of *telepathy*, but of *precognition*'. 'Overwhelming' was, of course, the word that Turing used in 'Computing machinery and intelligence' to characterize the evidence for ESP. (Leavitt, 2017, p. 350)

However, it would seem that even Leavitt was unaware that Broad could have played a far more direct role through his being co-editor of *Mind*, for in *The Man Who Knew Too Much*, his biography of Turing, he wrote: "One wonders what the editors of that august scientific publication *Mind* made of this bizarre appeal to a pseudoscience as baseless, if not as pernicious, as the one on the altar of which Turing would soon be laid out, as a kind of experiment" (Leavitt, 2006, p. 78).

Unshrouding Turing

Several thinkers have argued that the Turing Test, while of historical relevance, must be laid to rest in the cabinet files of AI history. Margaret Boden writes that passing the Turing Test is not in fact a terribly important goal or the most appropriate way to judge the progress of AI (Boden, 2006b, p. 1354). Already back in 1971, Bernard Meltzer, first editor of the *Artificial Intelligence* journal, suggested "that the Turing Test be retired, having done its proper work in the political battle to establish artificial intelligence as a respectable scientific discipline" (cited in McCorduck, 2004, p. 262). John Brockman, when introducing his anthology of think pieces from leading intellectuals on the future of machine thought claimed that "it's time to honor Turing and other AI pioneers by giving them a well-deserved rest" (2015, p. xxvi). Mark Halpern paints the relationship between Turing's memory and the AI community with the uneasiness that links esteemed and obsolete parents and their embarrassed adolescent offspring, who would claim that Turing's ideas "are no longer the foundation of AI work, and his paper may safely be relegated to the shelf where unread classics gather dust, even while we are asked to pay its author the profoundest respect" (2006, p. 45). Whitby (in French, 2000, p. 116) organized in a droll way four life-stages that the test has supposedly traversed:

> 1950–1966: a source of inspiration for all concerned with AI
>
> 1966–1973: a distraction from some more promising avenues of AI research
>
> 1973–1990: by now a source of distraction mainly to philosophers, rather than AI workers
>
> 1990 onwards: consigned to history

So why, then, should we insist on the worth of looking closely at it once more? Chiefly, because—regardless of how it may still inform or not, in an applicable way, the efforts of the legions of coders

employed by the currently reigning hi-tech megacorporations, trudging away under the banners of big data, neural nets and deep learning—there is something in *CMI* that appeals to us all, as humans.[36] In order to get to it, we first need to offer a defense against some of the most frequent criticisms Turing's proposed test faces.

Shieber (2004, p. 148) has streamlined the logical structure of a widely used type of argument to counter the validity of Turing's proposal. Given a test *T* claimed to assess the presence of a property *P*, a machine is described that could pass *T* without possessing *P*. Shieber calls such machines "Wedges" (as they drive a wedge between the property and the test that should be able to detect it). Shieber further identifies "Sparks" as that which would be missing from the Wedge, causing it to lack *P*. These Wedges are usually amusing, such as Keith Gunderson's (1964) box of rocks that proves beyond a shadow of a doubt that rocks can imitate, by virtue of their succeeding in the toe-stepping game (in which a participant must stick a toe through a hole and then decide whether it has been stomped on by a human foot or merely squashed by a falling rock) and Ned Block's (1981) Aunt Bertha machine, basically a humongous lookup table containing all the possible replies that Block's aunt would give to the finite but unimaginably large set of queries that could be put to her in under an hour. Copeland (2004, p. 437) has traced this family of objections all the way back to Shannon and McCarthy (1956) in *Automata Studies*.

But the most infamous specimen in this group is John Searle's (1980) Chinese Room *Gedankenexperiment*, which seldom wanders too far from Turing's on bibliographies, curricula and reference lists. Succinctly put, Searle tries to argue that even if a computer program exhibits verbal behavior that seems outwardly sophisticated, all of that is intrinsically no more than empty symbol manipulation, with no real understanding going on anywhere. In order to do this, he asks us to imagine him locked inside a gigantic contraption, in which his only means of communication with the outside world are slips of paper that he receives and delivers through a slit in the machine. Now, those slips have symbols in them, but not any kind of writing that he can understand. It is an alien sort of script or at least, alien enough, because to Searle, Chinese writing "is just so many meaningless squiggles" (p. 418). However, inside his room Searle has access to "a set of rules for correlating" (p. 418) some of these formal symbols with others according solely to their shapes. So, after he receives a slip of paper with its indecipherable squiggles and squoggles it is just a matter of looking those characters up in his instructions and handing back the corresponding formal symbols.[37]

In describing Searle's sleight of hand, Daniel Dennett (1980) introduced the concept of *intuition pump*, that is, "a device for provoking a family of intuitions by producing variations on a basic thought experiment" (p. 428) to end up concluding that Searle "relies almost entirely on ill-gotten gains: favorable intuitions generated by misleadingly presented thought experiments" (p. 429). Hofstadter (1980) argued that when we envision the scenario dreamt up by Searle we may fall for his ploy, his emotional trickery. We automatically empathize with the miserable soul trapped inside the Chinese room, condemned to carry out this joyless, daunting and monotonous task:

> Now Searle asks you to identify with this poor slave of a human (he doesn't actually ask you to identify with him - he merely knows you will project yourself onto this person, and vicariously experience the indescribably boring nightmare of that hand simulation. (1980, p. 434)

Therefore, what is going on takes on a flavor of artificiality, a mechanical quality divorced from life, and sparkle and thought, and we fail to realize that we are looking at the process from an immensely slowed-down level of description: "any time some phenomenon is looked at on a scale a million times different from its familiar scale, it doesn't seem the same!" (Hofstadter, 1980, p. 434).

Both Hofstadter and Dennett seem to agree that not only Searle's, but many of the other criticisms of the Turing Test can be accounted for by a failure of the imagination; their proponents are simply not seeing in enough depth the enormous complexity that the test demands and the subtle avenues for inquiring that it offers:

> Knowing that there was ferocious resistance to the image that computing machinery might soon, or indeed, ever, think, Turing took pains to point out the remarkable generality of the probing allowed by his test, by presenting a pair of short sample dialogues in which it was shown how a skillful human interrogator might try to elicit odd and recondite knowledge, subtle judgments, and even emotional responses from the unknown "being". But most people remain skeptical about the Turing Test even after reading these dialogues, probably because they fear that they might be easily taken in by the wiles of a superficial machine. They do not appreciate how deeply and broadly the Turing Test potentially would allow them to probe. (Hofstadter, 1985, p. 489)

Microcosmic Philosophy

Much like Haeckel when referring ontogeny to phylogeny, Warren Sack affirmed that "AI criticism has largely been a recapitulation, in miniature, of old, existing debates between rationalists, empiricists, romanticists, phenomenologists, pragmatists, and a handful of other named positions in the discourse

of western philosophy" (1997, p. 2). From a philosophical point of view, Turing's operationalization throws us a lifejacket while navigating the treacherous waters that stretch between the Scylla of solipsism and the Charybdis of panpsychism. For Turing saw that when taken to its utmost logical extremes, the original question inevitably led to the other-minds problem:

> According to the most extreme form of this view [the Argument from Consciousness as formulated by Professor Jefferson] the only way by which one could be sure that a machine thinks is to be the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. 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. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe "A thinks but B does not" whilst B believes "B thinks but A does not." Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks. (Turing, 1950, p. 446)

I believe Turing was well aware of how, in a sense, we are always running the imitation game he proposed, whether we realize or not. He offers a handy escape chute from solipsism, while, as ever, Descartes' evil genie looms above us, the past nearly five hundred years of philosophy not having managed to exorcize it. I will not claim that Turing synthesized the ultimate antivenin to its bite, but his injunction for civility can help us tackle the dilemma of solipsism in a way that is truly grounded upon a substrate of vitality, like Dr. Johnson kicking the proverbial stone to counteract Berkeley's doctrines. For indeed, no more than in the case of the machine can we access the inner sanctum of the private thoughts of those beings that we happily and nay-automatically count as our fellows (that is, unless we are afflicted by some variety of psychopathy or another empathetic malady). We could paraphrase Turing's "it is usual to have the polite convention that everyone thinks" by "it is more convenient to conduct ourselves as if they were." A pragmatic truth if there ever was one.[38]

It is this experimental and open-minded drive to find out for ourselves, whether our counterpart (machine or human) is minded or not that should guide our willingness to attribute such a description. The placing of both participants in rooms removed from sight is extremely democratic, as is Turing's insistence that this must be done precisely so that the outer appearance of the machine should not impact our judgment of the value of what it has to say. By orchestrating the proof at the level of function, Turing's goal was "not to penalise the machine for its inability to shine in beauty competitions" (1950, p. 434). He added that "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" (1950, p. 434).

Not only would it be pointless, but between certain specific ranges of anthropomorphism, it could even be downright counterproductive. Turing anticipated that phenomenon of strangeness felt in the face of close-but-not-quite-there simulacra which would later become known as the Uncanny Valley (Mori, 1970), a zone at the positive end of the spectrum of human resemblance whose every member elicits a high creepiness factor. It is far preferable—from the point of view of the feelings that it would stir in humans—for a machine to appear less human-like than to be a denizen of the valley:

> 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. (Turing, c. 1951, p. 486)

But what's with that tendency to conceptualize machines as entirely alien to us? In his *A Coffeehouse Conversation on the Turing Test*, Hofstadter (1985, p. 506) presents us with a dialogue that illustrates this aversion felt towards visualizing ourselves as machines. Here's a small abridged segment to give a taste of it:

> CHRIS: I'm not totally convinced that a machine is all I am. I admit my concept of machines probably does suffer from anachronistic subconscious flavors, but I'm afraid I can't change such a deeply rooted sense in a flash.
>
> SANDY: Part of me balks at calling myself a machine. It *is* a bizarre thought that a feeling being like you or me might emerge from mere circuitry.

Hofstadter relishes—and makes us relish too—the irony that such statements should come from entities that amount to no more than strings of printed characters on a page.[39] In that sense, Chris and Sandy are in fact even less than a machine, or are machines only in the restricted sense in which William Carlos Williams says that "a poem is a small (or large) machine made out of words" (1969, p. 256). This dislike or even repulsion towards thoughts of our own possible mechanicalness was addressed with poetic flair by the materialist Julien Offray de La Mettrie over 260 years ago. But, as we noted with Sack at the opening of this section, given the recursively recapitulatory nature of AI discourse, it is to be expected for such positions to be forgotten only for then to lie dormant in wait of their rediscovery:

> To be a machine and to feel, to think and to be able to distinguish right from wrong, like blue from yellow - in a word to be born with intelligence and a sure instinct for morality and to be only an animal - are thus things which are no more contradictory than to be an ape or a parrot and to be able to give oneself pleasure. For since here we have an opportunity to say so, who would ever have guessed a priori that a drop of liquid ejaculated in mating would provoke such divine pleasure and that from it would be born a little creature that one day, given certain laws, would be able to enjoy the same delights? I believe thought to be so little incompatible with organised matter that it seems to be one of its properties, like electricity, motive power, impenetrability, extension, etc. (La Mettrie, 1750/1996, p. 35)

Many of the negative associations to machines may be traced back to the very name of the field. John McCarthy chose 'artificial intelligence' as a way to distinguish it from the more conservative name of 'automata theory' (Kline, 2011; McCorduck, 2004) but not everyone was thrilled by the choice. Newell and Simon didn't like it and machine learning pioneer Arthur Samuel claimed that "the word artificial makes you think there's something kind of phony about this or else it sounds like it's all artificial and there's nothing real about this work at all"[40] (cited in McCorduck, 2004, p. 115).

> [I]t's about time we stopped using the term *artificial* in AI altogether. What we really mean is "designed intelligence" (DI). In popular parlance, words like *artificial* and *machine* are used in contradistinction to *natural* and carry overtones of metallic robots, electronic circuits, and digital computers, as opposed to living, pulsing, thinking biological organisms. (Davies, 2015, p. 29)

Jerome Bruner invokes Zipf's law—i.e., that the length of a word is inversely correlated with its frequency, at least according to Bruner—to explain the ubiquitous adoption of the initials AI when speaking about artificial intelligence: it would be the mark of its popularity. But he also points to a more interesting reason. The initialism would be in vogue because it plays a crucial euphemistic role "either because there is an aura of obscenity about the artificialization of something so natural as intelligence […] or because AI is an abbreviation of what, in its full form, might seem an oxymoron (the liveliness of intelligence coupled with the flatness of artificiality)" (1990, p. 10). Is there any legitimacy to the incompatibility between both terms that Bruner points at? Excluding the artificial (made with art) from the phenomenon of intelligence because of the 'liveliness' of the latter appears to be a *petitio principii*, for it is precisely the nature of intelligence what we are trying to understand.

But as Bruno Latour has pointed out, there is already a heavy abstraction at play in demarcating the machine from the non-machine, and in particular, in posing a strong separation between us and them.

He speaks of quasimachines, in that they cannot be fully separated (nor can we) from their attending circumstances:

> We would make great progress in our conversation, our exploration, it seems to me, if we did not tried [sic] to pass an impossible Turing test. I am loaded with too many machines already to pass it as a naked human […] Why don't we try instead to understand why we like so much to cut out among the entities making up our voices, bodies, engines, cities, institutions something, that would behave mechanistically? Let us decide to call "quasimachines" the open entities and "machine" the rendering of some of their parts as behaving mechanistically […] When you look at a space vessel cruising through space with unerring precision it looks like a machine, but if you suddenly hear the famous little warning "Houston we have a problem!" then it becomes clear that the space vessel is a quasi-machine that never left the umbilical chord of the huge institution down there on earth. (Latour & Powers, 1998, p. 6)[41]

The Danger of Tacit Assumptions

Now that we have hopefully given some inkling of what the test is about and reviewed both attacks on and defenses of it, it is time to introduce my apprehension, mentioned at the outset of the essay. Rosas has stressed how different logical levels of observation must be distinguished when analyzing the scenario of the Turing Test, and has offered the crucial insight that, in the context of a Turing Test, the machine only begins to be one once it is unearthed as one (1992, p. 120). The Turing Test as read by us in Turing's paper, or in any other paper dealing with it (such as this one) is fundamentally distinct from any instance of a Turing or near-Turing Test that we could actually face in reality[42], for in the first case (when considering the issue theoretically) we are given an omniscient assurance (even if only tacit) that a Turing Test is underway.

Just like by the mere conjuring of the man in the Chinese room we instinctively project ourselves in the role of this suffering individual burdened with such a daunting task, when faced with an exposition of Turing's scenario, we also automatically and semi-unconsciously assign to ourselves the role of the test's organizer, not the examiner, and furthermore, simultaneously, we fail to consider that we are also assuming a parallel role, that of an omniscient entity judging the examiner's accuracy and sound judgment. We reserve little space to the *urgency* of the question, for we forget that—as readers of the thought experiment—we are *given* the true state of the system, that is, the nature of the contenders. We should be on guard to not let that tacit confident attitude tag along when we are facing Turing Tests in real life.

What intuitive assumption is at play then when we analyze the Turing Test? For one, that we are already in possession of the notion that a computer IS being tested. We identify ourselves with those who design, or at least those who run, the experiment. It is difficult to take it seriously and it may even seem banal, for ultimately it comes down to whether a specific set of human judges are easily fooled or not.[43] But we should be more charitable toward it, since we ought to place ourselves in the position (we always are in the position!) of the examiner: we really have no idea what is behind either of those doors and we should thus approach them with the utmost open-mindedness we are capable of.

As a rudimentary way of exploring how such an intuitive perspective-taking could be modified, we can imagine what would happen if in our imagined scenario we significantly raised the stakes. Distinguishing the venomous coral snake from the outwardly similar but harmless scarlet kingsnake is a trivial concern when filling in a zoology written exam, but one of portentous consequences if, facing the two, someone forced us to grab one of them. Suddenly, it would appear as too steep a price to gamble away our life on having accurately memorized any of the trite nursery rhymes that are taught to American schoolchildren so that they will be able to distinguish them according to the arrangement of their colored bands (e.g. "red on yellow, kill a fellow; red on black, venom lack").[44]

For instance, let us suppose that there are two cabins in a spaceship, one containing a human passenger and one containing a psychopathic robot, and due to some technical malfunction they have both become locked and oxygen supply to one of them will cease after an interval sufficient for conducting a Turing Test. You, the ship's captain, can allocate the sole oxygen flow to either of the cabins, but must first ascertain whether the messages you are receiving come from your fellow human or from the murderous robot who, though untroubled by aerobic needs is simply trying to deprive the human of his needful oxygen out of pure spite. The sudden importance of making the correct identification (in this particular case, not 'who is a thinking being' but 'who is the human') should help align our perspective with that of the Turing examiner.

Self-directed vs. Mandated Sense-making

The enemy of our adopting and internalizing this attitude is, of course, the label. Labels are, to be sure, indispensable in our mental repertoire, but certain labels are also instrumental in stopping thought in its tracks. Of course, thought-stopping labels change with every age and season, but what is common to them all is that they provide a quick, easy and dirty release to the exertion of considering a complex problem. Whether the specific label that conveys a "seek no further" be actually called *mad*,

heretical, *child*, *senile* or *racist* is merely a contingency of the here and the now. I will call these labels that concern me 'overriding labels'. *Prima facie*, all labels would seem to disincline us to re-examine them, that is what they are there for. The expiration date on a package of Prosciutto is very seldom an invitation to perform an analysis of our own to ratify its verdict. Inevitably, then, any label must rest on some authority. What sets overriding labels apart and makes them worthy of being singled out, however, is that they pertain and become attached to beings who are in effect able to "read" them and either agree with them or not, and furthermore, explain why that is so. Yet the overriding label can forestall *us* from taking into account in the slightest the input from whatever has been thus labeled. We will not eat the once appetizing slices of dry-cured ham if they smell foul, no matter how compelling the credentials of the health board behind its best-by date. Why should we not exercise a similar modicum of self-directed judgment when dealing with something far more important?

Labels that seem to have been decreed by authority and are therefore imbued with its lingering touch bring the peril of offering to some a tempting allowance or license to not think by themselves. The idea that there is some sort of infallible authority we can turn to, that can provide such answers, is already an illusion. When faced with Turing's scenario we are neither the experimenters nor the readers, we are always the examiner: he who must decide, and face the consequences of such a decision. That is, precisely, at the core of Turing's argument and therefore why we have devoted so much attention to it. There is a deep integrity in finding one's own answers to one's own questions. Whoever forfeits this right and privilege, has allowed others to take in their hands the reins of his thought. The search for truth is a matter not only of scientific but also of personal integrity.

A chilling exploration of the implications of overriding labels was provided by David Rosenhan's (1973) *On Being Sane in Insane Places*. By having sane persons manage to be admitted into several psychiatric hospitals (they claimed falsely that they had been hearing voices), and then observing whether they would be detected as sane, Rosenhan attempted to show how once assigned, a marker of insanity is almost impossible to shed.[45] His vivid description of the ordeals faced by his pseudopatients plunges us emotionally into the implications of being unheard and being assigned a tag that cannot be defied. That is, of course, what would happen to thinking machines—if there ever were some such—if rather than guiding our assessment by Turing's operational criterion we were to allow ourselves to be guided by the external ascription of presumed preexistent properties. But machines are far from unique in this regard. Children and the very old are routinely denied the rights that most of the rest of us freely possess. Many of their decisions are not theirs to make, and the logic

behind this restriction is self-protecting. They have been assigned to the group of those who are not to determine the course of their lives.

The motif of trying to speak and not being heard plays a central role in Mary Shelley's (1818) *Frankenstein*. Frankenstein's monster is at "birth" endowed with a sound moral compass and a value system that is initially aligned with that of its human creator, but ends up developing a sociopathic streak, after several passages that are agonizing to read, as a result of feeling mistreated, excluded and misunderstood. He turns from humankind after being hunted down and attacked by those he was only trying to help, but who never gave him the tiniest chance to express himself. We appreciate a similar conflict played out in the contemporary ecological discussion over animal rights and animal sentience (e.g., Chapman & Huffman, 2018), where the lack of vocal language is argued to be an unfair ground upon which to impose pain upon feeling creatures. As primatologist Frans de Waal explains: "Language acquisition by animals became a huge topic that drew enormous public interest. It was as if all questions about animal intelligence boiled down to a sort of Turing test: can we, humans, hold a sensible conversation with them?" (de Waal, 2016, p. 54).

What we are concerned with here, is ultimately coming to grips with an almost by definition insurmountable obstacle; understanding alien intelligence. But this is what literature routinely does, and what hard sci-fi attempts to do under particularly strenuous conditions, and it seems also to be an innate mode of empathetic being in relation to others, when, to paraphrase Theodor Lipps' (1903) archetypal example, we feel ourselves into the acrobat when we see him up in the tightrope.[46]

We need not venture as far as the hallowed pages of classic literary fiction, however, for a good illustration of the same principle. A real case, albeit somewhat milder, lies at hand. The tragic story of Joseph Merrick, known as the Elephant Man, has been immortalized in film and the stage. His unusual deformity made him an outcast and the target of brutal scorn, mockery and ostracism. In a letter to the *Times*, seeking public assistance, Francis Carr-Gomm, chairman of London Hospital, shared his misery with the world:

> Terrible though his appearance is, so terrible indeed that women and nervous persons fly in terror from the sight of him, and that he is debarred from seeking to earn his livelihood in an ordinary way, yet he is superior in intelligence, can read and write, is quiet, gentle, not to say even refined in his mind. (Carr-Gomm, 1886/2014, p. 49)

And as if to prove that last point, slightly adapting a stanza from Isaac Watts[47], which he had probably learned during his childhood and once included in an autobiographical pamphlet, he penned the following little poem (Howell & Ford, 1983, p. 101) to accompany his words of thanks to the readers that had been moved to generosity by Mr. Carr-Gomm's missive:

Tis true my form is something odd,
But blaming me is blaming God;
Could I create myself anew
I would not fail in pleasing you.
If I could reach from pole to pole
Or grasp the ocean with a span,
I would be measured by the soul;
The mind's the standard of the man

Recalling Turing's sarcastic dismissal of the Theological Objection, if the Almighty has freedom to confer a soul on an elephant if He sees fit, certainly more so has He freedom to confer one on an Elephant Man, of which these tender verses render him deserving, and we have not only the capacity to detect it but the moral duty to do so.

There's no denying that there is an epistemological risk involved in approaching Turing Test-like scenarios with as open a mind as we can muster. As biological creatures with an evolutionary history, we are far from rational agents and can be easily fooled, even at a visceral level, by things we understand should not affect us[48]. But even that risk of erring in our judgement is to be preferred to the possibility of depriving a thinking agent of its social rights.[49] The best course of action, even if we can't avoid the occasional misstep along the way, is following Hofstadter's advice:

> Minds exist in brains and may come to exist in programmed machines. If and when such machines come about, their causal powers will derive not from the substances they are made of, but from their design and the programs that run in them. And the way we will know they have those causal powers is by talking to them and listening carefully to what they have to say. (1981/2000, p. 382)

No Machine Left Behind

In his book *Metamagical Themas*, Douglas Hofstadter recounts being invited to interact with Nicolai, a natural language program, by Zamir Babel, Professor of Computer Science at the University of Kansas and his students. What ensued was about an hour of back-and-forth conversation in which Nicolai alternated between dull rigidity and surprising bouts of seeming creativity. Finally, Nicolai

proved to be a bit just too clever for what Hofstadter was willing to believe and he suspected something odd was going on. It turned out that there was no Nicolai program, but just three students who were cleverly simulating being one (mechanical quirks included) through a remote terminal. This is a terrific tale (and, needless to say, far better told in the original) but what really fascinated (and deeply worried) me was the following passage near the end:

> It seems that a few days earlier, the class had collectively gone through something similar to what I had just gone through, with one major difference. Howard Darsche, who had impersonated (if I may use that peculiar choice of words!) Nicolai in the first run-through, simply had acted himself, without trying to feign mechanicalness in any way. When asked what color the sky was, he replied, "In daylight or at night?" and when told "At night", he replied, "Dark purple with stars." He got increasingly poetic and creative in his responses to the class, but no one grew suspicious that this Nicolai was a fraud. […] Zamir summarizes this dramatic demonstration by saying that *his class was willing to view anything on a video terminal as mechanically produced, no matter how sophisticated, insightful, or poetic an utterance it might be. They might find it interesting and even surprising, but they would find some way to discount those qualities*. (Hofstadter, 1985, p. 522, emphases added)

What Hofstadter's anecdote shows us is that there is always room for uncertainty, even if we have sound grounds to presume that there are tricks involved. Of even greater concern, that the students were able to disregard thinking beings as non-thinking merely because they sported a label associated with non-mentality (in this case, believing they were interacting with the computer program Nicolai). I find this terrifying. And it is not only unexperienced University students who are at risk of adopting such a frame of mind. Let us consider the following passage by Michael Polanyi:

> This apparent self-contradiction [between the authoritative pronouncements of science and its encouragement of creative dissent] is resolved on the metaphysical grounds which underlie all our knowledge of the external world. The sight of a solid object indicates that it has both another side and a hidden interior, which we could explore; the sight of another person points at unlimited hidden workings of his mind and body. Perception has this inexhaustible profundity, because what we perceive is an aspect of reality, and aspects of reality are clues to boundless undisclosed, and perhaps yet unthinkable, experiences. (Polanyi, 1966, p. 47)

This sounds all right, but it is already a double-edged knife, as we can see when we compare it to a previous remark of Polanyi's, made in his 1958 *Personal Knowledge*:

> [T]o acknowledge someone as a sane person is to establish a reciprocal relation to him. By virtue of our own art of comprehension we experience another person's similar faculties as the presence of that person's mind. (Polanyi, 1958, p. 277)

The real problem arises in what Polanyi says next:

> According to these definitions of 'mind' and 'person', neither a machine, nor a neurological model, nor an equivalent robot, can be said to think, feel, imagine, desire, mean, believe or judge something. They may conceivably simulate these propensities to such an extent as to deceive us altogether. But a deception, however compelling, does not qualify thereby as truth: no amount of subsequent experience can justify us in accepting as identical two things *known from the start to be different in their nature.* (Polanyi, 1958, p. 277, emphasis added)

This self-assured throwaway assertion by means of which Polanyi seems to settle the issue is precisely the kind of mental attitude that we have tried to battle all throughout this piece. It is that "known from the start to be different in their nature" that carries the seed of the danger of disregarding thinking beings as thinking. This is the very same thing that occurs when a machine's output is disregard as un-thinking and un-reflexive merely because it stems from a machine. That is why Turing's Test is so revolutionary: because it grants the benefit of the doubt and renders truthful that old decree; "by their fruits ye shall know them". Whether the machine "thinks" or not, cannot be known beforehand. It depends entirely upon what it tells us.

The crucial takeaway is that wrong or right answers do not preexist, but must be gained by each interrogator at the expense of a careful examination. Relying on a tag like 'machine' to conclude 'non-thinking' is a luxury we should be aware we cannot afford, lest we be left ethically impecunious. 'Machine' will in all likelihood be a tag imposed by a third party authority; we have a moral duty to personally decide, each time, whether it fits. Increasingly, as the technology progresses it will be harder and harder, in all likelihood completely impossible, for a single human mind to encompass all the complexity of the programming that will be involved in machines with outward displays of apparent intelligence. Even in such cases in which there are excellent reasons for assuming at the outset that we are dealing with an *unthinking* machine, there is space for subtle and careful consideration. The story of Joseph Weizenbaum and Eliza is well known, so I will refer the essentials. Weizenbaum programmed a very simple conversation chatbot around the idea of it simulating a non-directive Rogerian therapist, that is, a character who would mainly reflect back at its conversational counterpart

whatever it was that they had said first, possibly seeking elaboration or clarification, but offering no material of its own (McCorduck, 2004). The main problem for Weizenbaum was to find that several people fell for Eliza's simple trick and engaged with it in conversations that felt meaningful: they were finally being listened to. That sophisticated humans would easily assign human qualities to a rudimentary syntactic trick disgusted him, and like Victor Frankenstein, he recoiled in horror from what he had created (Copeland, 1993, p. 14). This definitely turned Weizenbaum off to AI, after which he became one of the field's most vocal critics.

Perhaps Weizenbaum *knew* that Eliza couldn't think; he had written the code, after all. (Though as we saw in Hofstadter's anecdote, he could have been tricked and his Eliza replaced with something else.) The problem here is that *we* are certainly not Weizenbaum, and while we may be told about a program "oh, that's just plain ole' Eliza," we will have no way of knowing for sure. I have to speculate here, but even so, if Weizenbaum had suddenly seen his program display a performance orders of magnitude more sophisticated than that which by virtue of his programming it was supposed to be able to do, that feat should have been grounds enough for questioning his initial assumptions. Even though Donald Davidson is no friend to the Turing Test as an assessment of meaning (see Davidson, 2004; Lohse, 2015), we would do well to lubricate our every interaction with speech-capable machines with a little bit of Davidsonian charity. Even Davidson allows that he would rather change his mind and his assumptions confronted with new strong evidence than stick to his current beliefs:

> Of course we believe, with good reason, that only creatures with a certain biological make-up actually do think; but if my friend turned out (after all these years) to be made of silicon, I'd change my mind about what materials a person might be made of, not my judgment that he was a person. (Davidson, 2004, p. 79)

If we are not willing to challenge such known facts in the face of compelling new evidence, we may as well chant along with Shaw's (1914/1994) Professor Higgins, when talking about his Eliza, presumably a sentient being, that served as namesake for Weizenbaum's program:

> PICKERING (*in good-humored remonstrance*): Does it occur to you, Higgins, that the girl has some feelings?
>
> HIGGINS (*looking critically at her*): Oh no, I don't think so. Not any feelings that we need bother about. (*Cheerily*) Have you, Eliza?
>
> LIZA: I got my feelings same as anyone else.

Imitating Turing

Many commentators have indulged a perfectly understandable impulse of reading Turing's Test biographically. The poetic temptation of seeing Turing as trying to pass his own sort of imitation test in settings in which he did not quite fit is nigh irresistible. While some overstate somewhat their case (e.g., Cowen & Dawson (2009) arguing for Turing's supposed autism) it is easy to sympathize with their angle of approach. In a letter he wrote to his friend Norman Routledge in 1952 (Hodges, 2000, p. xvi), two years before ending his life, Turing worried, for tragic and very justified reasons, about irrelevant attributes being used to rule out machine thinking:

> 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*

What I have been so insistently repeating again and again, that Turing's rejection of the Argument for Consciousness provides us with an epistemic tool to escape solipsism, can also be seen in Turing's very central plea that we should give "fair play to the machine" (Turing, 1947, p. 394), to which several prominent scholars of Turing draw attention (Copeland, 2004, p. 469; Proudfoot, 2017, p. 299; Leavitt, 2017, p. 355; Hodges, 2000, p. 361):

> If a computer, on the basis of its written replies to questions, could not be distinguished from a human respondent, then 'fair play' would oblige one to say that it must be 'thinking'. This being a philosophical paper, he produced an argument in favour of adopting the imitation principle as a criterion. This was that there was no way of telling that other *people* were 'thinking' or 'conscious' except by a process of comparison with oneself, and he saw no reason to treat computers any differently. (Hodges, 2000, p. 415)

As both Hodges (2000) and Hofstadter (1979/2000) pointed out, Turing's willingness to examine or consider what seemed to be the substantial evidence for ESP instead of outright rejecting it because it was the fashionable thing to do, is a sign of his vast intellectual and scientific integrity. Conversely, C.D. Broad's earlier-quoted statement on the paranormal would make perfect sense if one imagined that he was referring instead to the possibility of machine intelligence: "A man must know a great deal more about the secrets of nature than I do to reject any alleged fact without investigation, however wild it may seem" (Broad, 1924, §3). That is precisely the defiant attitude adopted by Turing. At the

very beginning of his well-titled talk, *Intelligent Machinery, A Heretical Theory*, aired by the BBC radio, he stated: " '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, 1951, p. 472).

References

Abramson, D. (2011). Descartes' influence on Turing. *Studies in History and Philosophy of Science* Part A, 42(4), pp. 544–551.

Annan, N.G., Attlee, C.R., Ayer, A.J., Boothby, R., Bowra, C.M., Broad, C.D. … Wooton, B. (1958, March 7). Letter to the Editor. *The Times*.

Barrat, J. (2013). *Our Final Invention: Artificial Intelligence and the End of the Human Era*. Chicago: Thomas Dunne Books.

Block, N. (1990). The Computer Model of the Mind. In D.N. Osherson & H. Lasnik (Eds.), *An Invitation to Cognitive Science, vol. iii* Cambridge, Mass.: MIT Press.

Boring, E.G. (1946). Mind and Mechanism. *American Journal of Psychology*, Vol. 59, No. 2, pp. 173-192.

Broad, C.D. (1924). Critical and Speculative Philosophy. In J.H. Muirhead (Ed.) *Contemporary British Philosophy: Personal Statements*. London: G. Allen and Unwin. pp. 77-100. Available online at: http://www.ditext.com/broad/csp.html#b

Brockman, J. (2015). Preface: The 2015 Edge Question. In J. Brockman (Ed.), *What to Think About Machines That Think*. NY: HarperCollins. pp. xxv-xxvi

Bruner, J. (1990). *Acts of Meaning*. Cambridge, MA: Harvard University Press.

Burks, A. (2003). *Who Invented the Computer?: The Legal Battle that Changed Computing History*. Amherst, N.Y: Prometheus Books.

Calvino, I. (2000). *Why Read The Classics?* NY: Vintage.

Carr-Gomm, F. (2014). Letter to the Times. In S. Usher (Ed.) *Letters of Note: An Eclectic Collection of Correspondence Deserving of a Wider Audience*. San Francisco: Chronicle Books. (Original letter published December 4, 1886).

Chapman, C.A. & Huffman, M.A. (2018). Why do we want to think humans are different? *Animal Sentience* 23(1)

Christian, B. (2011). *The Most Human Human: What Artificial Intelligence Teaches Us About What it Means to Be Alive*. NY: Doubleday.

Copeland, B.J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

Copeland, B.J. (2004). *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press.

Cornejo, C. (2013). On Trust and Distrust in the Lifeworld. In P. Linell & I. Marková. (Eds.), *Dialogical Approaches to Trust in Communication*, pp. 237–254.

Cornejo, C. (2016). From Fantasy to Imagination: A Cultural History and a Moral for Cultural Psychology. In B. Wagoner, I. Bresco, & S.H. Awad, (Eds.) *The Psychology of Imagination: History, Theory and New Research Horizons*. Charlotte, NC: Information Age.

Cowen, T. & Dawson, M. (2009). *What Does the Turing Test Really Mean? And How Many Human Beings (Including Turing) Could Pass?* [Working Paper]. George Mason University. http://www.gmu.edu/centers/publicchoice/facultypages/Tyler/turingfinal.pdf

Davidson, D. (2004). *Problems of Rationality*. Oxford: Clarendon Press.

Davis, D.A. (1976). On being detectably sane in insane places: Base rates and psychodiagnosis. *Journal of Abnormal Psychology*, 85(4), pp. 416-422

Dennett, D. (1980). The milk of human intentionality. *Behavioral and Brain Sciences* 3, pp. 428–430.

Dennett, D. (2004). Can Machines Think? In C. Teuscher (Ed.) *Alan Turing: Life and Legacy of a Great Thinker*. Berlin: Springer-Verlag, pp. 295-316.

de Waal, F. (2016). *Are We Smart Enough to Know How Smart Animals Are?* NY: W.W. Norton & Company.

French, R.M. (1990). Subcognition and the limits of the Turing test. *Mind*, 99(393), pp. 53-65.

French, R.M. (2000). The Turing Test: The First 50 Years. *Trends in Cognitive Sciences, 4*, pp. 115–22.

French, R.M. (2012). Dusting Off the Turing Test. *Science*, 336(6078), pp. 164–165.

Gunderson, K. (1964). The imitation game. In A.R. Anderson (Ed.) *Mind and Machines*. (Englewood Cliffe, NJ: Prentice Hall.

Halpern, M. (2006). The trouble with the Turing test. *The New Atlantis*, 11, pp. 42–63.

Hodges, A. (1992). *Alan Turing: The Enigma*. London: Vintage.

Hodges, A. (1997). *Turing: A Natural Philosopher*. London: Phoenix.

Hodges, A. (2000). *Alan Turing: The Enigma*. NY: Walker.

Hofstadter, D. (1980). Reductionism and religion. *Behavioral and Brain Sciences* 3, pp. 433–434.

Hofstadter, D. (1985). *Metamagical Themas: Questing for the Essence of Mind and Pattern*. NY: Basic Books.

Hofstadter, D. (2000). *Gödel, Escher, Bach: An Eternal Golden Braid*. London: Penguin. (Original work published 1979).

Hofstadter, D. (2000). Reflections on 'Minds, Brains, and Programs'. In D. Hofstadter & D. Dennett (Eds.), *The Mind's I: Fantasies and Reflections on Self and Soul*. NY: Basic Books, pp. 373-382. (Original work published 1981).

Howell, M. & Ford, P. (1983). *The True History of the Elephant Man*. London: Allison & Busby.

James, W. (1890). *The Principles of Psychology*. New York: Henry Holt and Company.

James, W. (1922). *Pragmatism: A New Name for Some Old Ways of Thinking*. NY: Longmans, Green and Co.

Kesey, K. (1962). *One Flew Over the Cuckoo's Nest*. New York: Signet.

La Mettrie, J. O. de, (1996). *Machine Man and Other Writings* . [Translated and Edited by Ann Thomson]. Cambridge, UK: Cambridge University Press. (Original work published 1750).

Latour B., & Powers, R. (1998). Two writers facing one Turing Test: A dialog in honor of HAL between Richard Powers and Bruno Latour. *Common Knowledge* 7(1):177–191.

Leavitt, D. (2006). *The Man Who Knew Too Much: Alan Turing and the Invention of the Computer*. NY: W. W. Norton.

Leavitt, D. (2017). Turing and the paranormal. In B.J. Copeland, J. Bowen, M. Sprevak & R. Wilson (Eds.) *The Turing Guide*. Oxford: Oxford University Press, pp. 347-356.

Lipps, T. (1903). *Grundlegung der Ästhetik*. [Foundations of aesthetics]. Hamburg & Leipzig: Leopold Voss.

Loebner, H. (1994). In response. *Communications of the ACM* 37.6, pp. 79–82.

Lohse, T. (2015). *Davidson's Test: Donald Davidson's Critique of the Turing Test as an Expression of His Theory of Intellectual and Linguistic Competence*. (Dissertation). Retrieved from: http://mrloh.se/assets/doc/davidsons_test_TL_2015.pdf

McCorduck, P. (2004). *Machines Who Think*. Natick, Massachussets: A K Peters.

McPherson, T. (1950). The existence of God. *Mind*, 59 (236), 545–550.

Millon, T. (1975). Reflections on Rosenhan's "On being sane in insane places." *Journal of Abnormal Psychology*, 84(5), pp. 456-461.

Mori, M. (1970). The uncanny valley. *Energy* 7, 33–35.

Negrotti, M. (2002). *Naturoids: On the Nature of the Artificial*. Singapore: World Scientific Publishing.

Newman, J.R. (1956). *The World of Mathematics*. NY: Simon & Schuster.

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago: University of Chicago Press.

Polanyi, M. (1966). *The Tacit Dimension*. Garden City, N.Y.: Doubleday.

Proudfoot, D. (2017). The Turing test—from every angle. In B.J. Copeland, J. Bowen, M. Sprevak & R. Wilson (Eds.) *The Turing Guide*. Oxford: Oxford University Press, pp. 287-300.

Rosas, R. (1992). ¿Comerán los androides el fruto prohibido? Reflexiones acerca del Test de Turing. *Apuntes de Ingeniería* 45 (1992) pp. 111-129.

Rosenhan, D. (1973). On Being Sane in Insane Places. *Science*, 179(4070), pp. 250-258.

Rosenhan, D. (1975). The Contextual Nature of Psychiatric Diagnosis. *Journal of Abnormal Psychology*, 84(5), pp. 462-474.

Sack, W. (1996). *Replaying Turing's Imitation Game*. Paper resented at the panel "Nets and Internets" at Console-ing Passions: Television, Video and Feminism, April 25-28, 1996, Madison, WI. Available at: http://www.pd.org/Perforations/perf13/wsachs.html

Sack, W. (1997). Artificial Human Nature. *Design Issues*, 13. pp. 55-64.

Searle, J. R. (1980) Minds, brains and programs. *Behavioral and Brain Sciences* 3, pp. 417-424.

Seneca, L.A. (2015). *Letters on Ethics To Lucilius*. [Translated and edited by M. Graver and A. A. Long]. Chicago: University of Chicago Press.

Shaw, G.B. (1994). *Pygmalion*. Mineola, NY: Dover. (Original work published 1914).

Shannon, C. & McCarthy J. (1956). *Automata Studies*. Princeton: Princeton University Press.

Shelley, M. (1818). *Frankenstein; or, the Modern Prometheus*. London: M. K. Joseph.

Shieber, S. (1994). Lessons from a restricted Turing test. *Communications of the ACM* 37.6. pp. 70- 78

Shieber, S. (2004). The Wedge and the Spark. In S. Shieber (Ed.) *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, Mass: MIT Press, pp. 147-150.

Spitzer, R. (1975). On pseudoscience in science, logic in remission, and psychiatric diagnosis: A critique of Rosenhan's "On being sane in insane places". *Journal of Abnormal Psychology*, 84(5), pp. 442-452.

Spitzer, R. (1976) 'More on pseudoscience in science and the case for psychiatric diagnosis', *Archives of General Psychiatry*, 33: 459.

Sundman, J. (2003). Artificial stupidity. *Salon*. Retrieved from: https://web.archive.org/web/20120720014628/http://www.salon.com/2003/02/26/loebner_part_one

Turing, A.M. (1947). Lecture on the Automatic Computing Engine. In B.J. Copeland (Ed.) *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press, pp. 378-394.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind*, 59 (236), pp. 433–460.

Turing, A.M. (c. 1951). Intelligent Machinery: An Heretical Theory. In B.J. Copeland (Ed.) *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press, pp. 472-475.

Turing, A.M. (1951). Can Digital Computers Think? In B.J. Copeland (Ed.) *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press, pp. 482-486.

Watts, I. (1762). *Horae Lyricae: Poems, Chiefly of the Lyric Kind*. NY: Hugh Gaine, bookseller.

Williams, W. C. (1969). *Selected Essays of William Carlos Williams*. NY: New Directions.

Gamifying Programs

Roberto Musa G.

"Machine technology remains up to now
the most visible outgrowth of the essence of modern technology,
which is identical with the essence of modern metaphysics."

Martin Heidegger, 1977, p. 116

Abstract

Play has been crucial in the evolution of our species and has held a most important role in the development of our electronic scions as well. As the boundaries between human and machine become increasingly blurred, so do those between work and play and between life and games. The imminent advent of virtual reality as the favored landscape for human interaction, and the zest with which gamification has inherited the aims and strategies of prescriptive psychology are examples of how games will become a ferocious shaping force in the coming century. Current trends lead us to believe that there looms a non-negligible chance that intelligent programs, once reared in the playing of games like Chess and Checkers in their infancy, shall now transition to provide the games that will assuage an infantilized humankind.

Keywords: Artificial Intelligence, gamification, virtual reality, computer chess

It hardly needs stating how multithreaded the fabric of human experience is. Almost every aspect of our collective or individual, cultural or biological lives (and upon scrutiny the precise boundaries of each of those demarcations cannot fail but to dissolve into a field of rich and fuzzy fractal blurriness) can be unspun into gripping narratives and be visited as a treasure house of insights for and into the mind. But just like the edifice of what being human is all about offers countless doors in but no Grand Unified Theory, the same restriction is in effect when dealing with one of the latest offshoots of our collective ingenuity: Artificial Intelligence. Speaking about what our thinking machines are, unavoidably requires talking in turn about what we are. (And when this is not done explicitly, tacit assumptions seriously risk muddling our conversations). In this piece, I have chosen games and play as an entry point for a discussion about the relationship between humankind and its machines which today show promise for eventual autonomous thought. I hope to show that games are not only a lavishly detailed telescope both into the development of our species and into the progressive evolution of our creations, but, more importantly for our present purposes, that they also afford us the chance to groundedly anticipate how our mutual interactions may continue to unfold. Games were a crucial stepping stone in the development of mechanical thinking systems and were it not for them, the state of advance that we see now and take for granted would in all likelihood have taken considerably longer to achieve. And if our machines learn to bypass us and play among themselves, what mind-boggling outcomes could we reasonably expect down the road? We are already seeing notable signs of the power of adversarial self-play for artificial virtual agents, with researchers at OpenAI having recently shown how emergent strategies of high complexity arise without direct human instruction in a simulated game of hide-and-seek; strategies some of which even the researchers themselves had not known were possible within their system (Baker et al., 2019). In what follows I shall chiefly focus on two possible outcomes for our relationships with thinking machines, mutual collaboration and merging or utter dependency (a third and darker possibility, annihilation and replacement is not mentioned at length here but is explored in [MYTHS]).

At the outset, I plead the reader be indulgent with my having borrowed the circular scaffolding of my title's structure from Heinz von Foerster's *Observing Systems* (or alternately, Jonathan Safran Foer's *Eating Animals*). The choice boils down not to the mere appeal (always present and always tempting) of cutesy, low-hanging "clever" wording, but is warranted by the main thrust of the argument. The term Gamifying Programs is valid at both levels in which it can be interpreted. Stated laconically, it works as a gerund because we (as the *Homo sapiens sapiens* species) have gamified the environments of our

programs and it also works as a participle because it is quite plausible that programs may eventually gamify our reality. But before we plunge into gamification proper, and offer some working definition thereof, let us first zoom back and speak more generally about games, how they are inextricably linked with what it means to be human, and how they have shaped the history of AI.

Humans at play

There comes a point in the life of every psychologist where he or she must—or so at least suggests Daniel Gilbert—complete what Gilbert has ominously dubbed *The Sentence,* that is, provide the missing term in the fundamental statement: *The human being is the only animal that* _____. (2006, p. 10). While pet owners everywhere would be off-put at the short-sightedness of ending the sentence with 'plays', there is a sense in which human play is distinctively unique and a fundamental shaping force in our culture and history. Many different species play too, but the immense variety of our repertoire of play and its centrality in our lives is staggering. As W. Grey Walter, father of the electro-mechanical tortoises considered to be the first biologically inspired robots (Holland, 2003), expressed it: "It has been suggested that the greater an animal's brain the more its survival depends on the nature of its play. Human society devotes an enormous proportion of time and energy to play" (Walter, 1961, p. 225).

The cornerstone role of play in human culture has been expounded in depth by scholars the likes of Johan Huizinga (1980) and Roger Caillois (2001), in works of a caliber that render any attempt at cursory summarization a disservice to the scope of their journeying. Huizinga may be even considered to have anticipated how fundamental play was in the emergence of mind (Contreras, 2019, p. 29). And what is play? Saint Augustine's strategy regarding time ("If no one asks me, I know : if I wish to explain it to one that asketh, I know not") (1876, p. 235) and Justice Stewart's regarding hardcore pornography ("I know it when I see it") (*Jacobellis v. Ohio*, 1964) have been a source of succour for many attempting to survey expansive conceptual landscapes. As Gordon Burghardt put it in his seminal analysis of animal play, "when trying to sort out the boundaries of play, one quickly gets tangled in a web of definitions, controversies, and elusive notions that slip away just when one thinks that they are grasped." (2005, p. xi). Huizinga himself cautioned against adopting the alluring stance of deeming all human activity 'play', saying that such a move was "ancient wisdom, but it is also a little cheap" (1980, p. ix). Nevertheless, and fully acknowledging the slipperiness of the endeavor, let us take a part of Huizinga's definition of play as a starting point:

> [A] free activity standing quite consciously outside "ordinary" life as being "not serious," but at the same time absorbing the player intensely and utterly. It is an activity connected with no material interest, and no profit can be gained by it. It proceeds within its own proper boundaries of time and space according to fixed rules and in an orderly manner. (Huizinga, 1980, p. 13)

Commenting on it, Caillois observed that Huizinga's definition, "in which all the words are important and meaningful, is at the same time too broad and too narrow." (2001, p. 4). James P. Carse, in his thought-provoking and much revisitable *Finite and Infinite Games* (1986), a boundary-defying work (as evidenced by its subtitle 'A Vision of Life as Play and Possibility'), lays special stress in one of the dimensions of the definition; that of always being undertook freely and voluntarily: "It is an invariable principle of all play, finite and infinite, that whoever plays, plays freely. Whoever *must* play, cannot *play*" (p. 4, emphases in the original). Of these defining features of play highlighted by Huizinga, three shall occupy us especially in a latter section—its being "free", "outside life" and "unprofitable"—for as we shall ponder, their status as necessary traits may possibly be brought into question by the reality of contemporary video games and the social fabric in which they are embedded. What ultimately cannot be denied is that play is fundamental for us, and to an even greater extent than it has been for our phylogenic precursors. But what about our machines?

Early Game

When did computer programs undertake the first steps on the road to becoming our playmates? In her chronicle of the dawn of AI, Pamela McCorduck (1979) allotted a full chapter to the impact of games in the early days of the enterprise, which fittingly begins by subscribing Huizinga's proposed taxonomic label for our species: *Homo ludens*. To her it is no surprise that computers were involved in games from their earliest inception. Most of the early researchers in AI were involved with games, be it in their research or as a hobby (Skinner, 2016, p. 29). McCorduck ponders different explanations for why this should be the case. On the one hand, games serve as microdomains (an idea that we see again and again in the literature); they are simplified models of situations in real life, which express their essence, just like physical models imitate physical reality (p. 147). On the other hand, McCorduck considers, quoting the influential early anthology on AI edited by Feigenbaum and Feldman, "it provides a direct contest between man's wit and machine's wit" (cited in McCorduck, 1979, p. 147). But ultimately, she concludes that the true reason is not to be found in either of those façades, but rather in that "games are deep in the heart of us" (p. 146). "I've seen too many gleaming eyes to believe otherwise" (p. 148), she tells us.

Yet the importance of games as toy models of reality is not to be disregarded. The virtues of studying processes in restricted domains as opposed to immensely complex ones such as the entire physical universe or a fully functioning human brain become apparent in the following simile offered by Douglas Hofstadter (which, it must be clarified, was drawn not in relation to games, but rather, to a very special analogy-making program, Copycat, developed by him and Melanie Mitchell). This is therefore an analogy to an analogy between analogy-making programs, but I feel it is suggestive of the pros and cons of microdomains in general:

> Suppose one wanted to create an exhibit explaining the nature of feline life to an intelligent alien creature made of, say, interstellar plasma or some substrate radically different from that of terrestrial life. The Copycat approach might be likened to the strategy of sending a live ant along with some commentary aimed at relating this rather simple creature to its far larger, far more complex feline cousins. The rival approach might be likened to the strategy of sending along a battery-operated stuffed animal—a cute and furry life-sized toy kitty that could meow and purr and walk. This strategy preserves the surface-level size and appearance of cats, as well as some rudimentary actions, while sacrificing faithfulness to the deep processes of life itself, whereas the previous strategy, sacrificing nearly all surface appearances, concentrates instead on conveying the abstract processes of life in a tiny example and attempts to remedy that example's defects by explicitly describing some of what changes when you scale up the model. (Hofstadter, 1995, p. 302)

And among all AI toy models, none has been as fruitful as computer chess, the history of which is peppered with fascinating detail. Credit as the creator of the first artificial chess player could either go (depending on the stringency of our requirements to consider them as such) to Leonardo Torres y Quevedo, for his 1912 automaton, capable of playing a rook and king endgame, or to Dietrich Prinz, a colleague of Turing's, whose programming an electronic computer to play chess for the first time "was akin to the Wright brothers' first short flight. He had shown that computers were not just high-speed number crunchers. A computer had played chess" (Copeland, 2017, p. 342). However, there is another creation which, while not being a full-fledged member of the category, certainly deserves mention in any story of AI chess: Baron Von Kempelen's Turk.

This wondrous contraption—which artfully allowed a hidden human chess master to control a fancifully clothed mannequin which moved the chess pieces—toured Europe during the late eighteenth and early nineteenth centuries (first with its creator, Von Kempelen, and later with impresario Johann Maelzel) and eventually made its way to the United States. The reason for

summoning this odd character out of the history books is that the Turk is the perfect metaphor for the hidden human component often concealed in automation that is touted as independent. It is not surprising for Amazon to have called its crowdsourcing platform for 'human intelligence tasks' *Mechanical Turk*. As we shall consider in more depth in [OUTRO] there is a hidden layer of human intelligence in the seemingly clever displays of computer programs, which nevertheless can shine through the cracks when we look more carefully. The exploitation of fossilized human Big Data confronts us with the unescapable Lovelace Objection to machine originality (see Turing, 1950, p. 450; du Sautoy, 2019).

The Turk played against some of the most important figures of the age, like Benjamin Franklin and Catherine the Great (Standage, 2002a), and even defeated Napoleon, who trying to test the machine, went as far as to attempt some illegal moves during the game (Levy & Newborn, 2012). "Napoleon was better versed in the art of manoeuvring human kings, queens, prelates and pawns on the great chess-boards of diplomacy and battle than moving ivory chessmen on a painted table-top" concludes Henry Ridgely Evans in his essay, *The Romance of Automata* (1906, p. 135). Far more important for our purposes, however, was the encounter between the Turk and Charles Babbage, who twice defied it and twice lost (Standage, 2002b). While he suspected that the machine's performance was merely the trickery of a concealed human (as Edgar Allan Poe also later would[50]), the encounter seems to have left a lasting influence on his thoughts regarding the potential mental capabilities of machines (Standage, 2002b). Competent chess playing seemed to be a perfect benchmark of what only human-level intelligence could accomplish, which made building a genuinely autonomous chess playing machine so enticing a prospect as it remained for decades. As Babbage put it in his autobiography: "I endeavoured to ascertain the opinions of persons in every class of life and of all ages, whether they thought it required human reason to play games of skill. The almost constant answer was in the affirmative" (Babbage, 1864, p 465).

Babbage did not succeed in his ambition, and for years chess was heralded as a grail of human intelligence upon which machines would not trespass. In his scathing attack on the field, *Alchemy and Artificial Intelligence*, Hubert Dreyfus (1965) chose the very limited progress that machines were making on the chessboard as a clear sign of stagnation. Only a year later, his fiercest critic, Seymour Papert, arranged for Dreyfus to play against Richard Greenblatt's MacHack program (Boden, 2006, p. 841) where he "had the pleasure of […] seeing him very roundly trounced" (Papert, 1968, p. I-6). And yet, as important a criterion as human-level chess playing was held to be, once attained, it too fell prey to

what is commonly termed the *AI Effect*, which states that once AI becomes able to do something, then such a thing is no longer thought a hallmark of intelligence.[51] The same thing would later happen with the game of Go (du Sautoy, 2019).

Far from being only a pastime, chess has served as "a test-bed for ideas in Artificial Intelligence" (Copeland, 2004, p. 562) and has even been considered the 'standard organism' of AI (Ekbia, 2008; Ensmenger, 2011). Donald Michie, Turing's close collaborator and a crucial evangelist for his ideas, spreading them in several AI labs and universities in the UK and North America (Copeland, 2017, p. 267) drew the analogy explicitly:

> Computer chess has been described as the *Drosophila melanogaster* of machine intelligence. Just as Thomas Hunt Morgan and his colleagues were able to exploit the special limitations and conveniences of the *Drosophila* fruit fly to develop a methodology of genetic mapping, so the game of chess holds special interest for the study of the representation of human knowledge in machines. (cited in Copeland, 2004, p. 562)

Two names stand out especially among the many AI pioneers who set their sights on the problem: Claude Shannon and Alan Turing. Shannon, father of information theory,[52] wrote an influential paper outlining a computing routine that would enable a modern general purpose computer to play chess (Shannon, 1950, p. 256). Shannon seemed to have been aware that games are often thought frivolities, for in a latter write-up of his chess ideas, he gave a defense of the general utility that insight generated dealing with chess could provide for other areas:

> This problem, of course, is of no importance in itself, but it was undertaken with a serious purpose in mind. The investigation of the chess-playing problem is intended to develop techniques that can be used for more practical applications. (Shannon, 1956, p. 2124)

As thorough chronicler of the history of cybernetics Ronald Kline has documented, Shannon sympathized with the main goal of attaining human level artificial intelligence. In a letter to a former teacher, he declared that his fondest dream was "to someday build a machine that really thinks, learns, communicates with humans and manipulates its environment in a fairly sophisticated way" (Kline, 2011, p. 8).

Turing's thinking about computer chess was deeply at play in his reflections on whether intelligence could be mechanized. In 1948 he had collaborated with his friend, the statistician David

Champernowne, to produce the rules for a chess playing paper machine, affectionately given the moniker Turochamp after its creators (Copeland, 2017, p. 331). Turing actually started coding a revised version of this chess engine in the Manchester computer but never completed it. Interestingly, both Turing's and Shannon's chess engines have been contemporarily instantiated[53] in actual software (Copeland, 2017, p. 344) and made to compete against one another, the result being a tie after ten games (each program winning once and coming to a draw in the remaining eight). B. Jack Copeland, unabashed torchbearer of Turing that he is, however, is compelled to add that "given that repetitive moves often cost the Turing Engine its win, it seems probable that Turing would have beaten Shannon hands down had [a repetition detection] rule been in place" (p. 345).

We can further appreciate the theoretical boon that game-playing machines bestowed when we note that, in seed form, Turing's ideas that would later lead him to flesh out his now ubiquitous test in *Computing Machinery and Intelligence* (Turing, 1950; see [TESTS] for a detailed discussion of the article and its implications) had made a previous appearance in connection to his discussing chess playing:

> 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. […] 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 [a human operator precisely following the rules of an algorithm] 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. (Turing, 1948, p. 431)

Sadly, this article—which Copeland considers "effectively the first manifesto of AI" (2004, p. 355)—never saw the light of day, owing quite possibly to its negative reception on the part of Turing's superior at the National Physical Laboratory, where he had started working after the war. Oddly enough, the man in question was Charles Galton Darwin, grandson of Charles Darwin and godson of Francis Galton. The "headmasterly" C.G. Darwin, as Copeland puts it, deemed Turing's manifesto a "schoolboy's essay" (2004, p. 401) and argued against publication. Despite being connected both by nature and nurture to two of the most vivaciously inquiring spirits England ever produced, he himself failed to display the flight of scientific imagination needed to value the far-reaching implications of

Turing's precocious vision. In his gloomy speculative treatise predicting "the next million years" of humanity Darwin, barely spares a word for the "new high-speed calculating machines" relegating them to the possible role of uncannily accurate forecasters of competing policies, a task which they could undertake "with a completeness that is far beyond anything that the human mind can aspire to achieve directly" (Darwin, 1952, p. 55).

Unbeatable

But then again, we should perhaps not be too harsh on Darwin for failing to predict what was to come, for he is far from alone in that all too human failing. It is with a gasping pang of mute dread that we must oftentimes confront the archival remains of what accounts of the future the past dreamt up. Wisława Szymborska (1981, p. 121) has captured the feeling superbly in her poem *The Letters of the Dead*:

> *We read the letters of the dead like helpless gods,*
> *yet gods for all that, since we know the dates to come.*
> …
> *We silently observe their pawns on the chessboard,*
> *except they're now moved three squares further.*
> *Everything they foresaw came out quite different.*

We get a similar sensation when reading in what directions Artificial Intelligence pioneers thought the field would evolve. Here is Donald Michie in 1972 with his predictions on the (by then) future of computer chess:

> Hence if the knowledge of the chess-master were built into a computer program we would see not master chess, but something very much stronger. As with other sectors of machine intelligence, rich rewards await even partial solutions to the representation problem. To capture in a formal descriptive scheme the game's delicate structure—it is here that future progress lies, rather than in nanosecond access times, parallel processing, or mega-mega-bit memories. An interesting possibility which arises from the "brute force" capabilities of contemporary chess programs is the introduction of a new brand of "consultation chess" where the partnership is between man and machine. The human player would use the program to do extensive and tricky forward analyses of variations selected by his own chess knowledge and intuition, and to check out proposed lines of play for hidden flaws. (Michie, 1972, p. 332)

Whether we like it or not, in the end it was brute force that did the trick. We skip over the improvements that later decades brought, during which confident predictions time and again had to be readjusted. But finally, the frequently rescheduled promises of computer chess found their most iconic fulfillment in the victory of IBM's DeepBlue over Garry Kasparov in 1997: A computer program had beaten the reigning world chess champion. Unfortunately, little light was shed on the actual mental processes that underlie how humans think when engaged in the practice of chess, as the original pursuers of machine chess had hoped:

> The huge improvement in computer chess since Turing's day owes much more to advances in hardware engineering than to advances in AI. Massive increases in cpu speed and memory have meant that successive generations of machines have been able to examine increasingly many possible moves. Turing's expectation was that chess-programming would contribute to the study of how human beings think. In fact, little or nothing about human thought processes appears to have been learned from the series of projects that culminated in Deep Blue. (Copeland, 2004, p. 566)

Some, however, hope to learn "what computer-generated gameplay suggests about how brains operate" in the workings of the currently reigning and vastly superior AlphaZero (Purves, 2019, p. 14785). This deep neural network would show that "algorithmic computation (executing a series of specified steps)" (p. 14787) must be replaced as an analogy for how humans think in favor of "connectivity generated by trial-and-error learning over evolutionary and individual time" (p. 14786). It is important to mention (a fact that Purves acknowledges, but to an important extent downplays) that AlphaZero's approach of trial and error is not fundamentally new. Of course, it is not strange for past technological developments to be dropped or minimized in retellings[54], but the crux of what makes AlphaZero tick had already been thought of and partially developed in the 50s; we just lacked the hardware for it to work on the scale it now does:

> The learning procedure that Turing proposes in 'Chess' involves the machine trying out variations in its method of play—e.g. varying the numerical values that are assigned to the various pieces. The machine adopts any variation that leads to more satisfactory results. This procedure is an early example of a genetic algorithm. (Copeland, 2004, p. 565)[55]

More important still, is the fact that there is no such clear-cut distinction between "rule-based computation" (Purves, 2019, p. 14787) and "learning on a wholly empirical (trial and error) basis" (p. 14786). As Esteban Hurtado has clarified in his work on the limitations of computer models for

human thought, "the usual way of implementing a neural network is by means of a programing language that uses the same old rigid formal rules. So, actually, neural networks, as a theoretical mind modeling device, do not add any new capability" (Hurtado, 2017, p. 3).

But even if AlphaZero will not show us a path to better understand human thought, it may be on the way to develop its own very distinct spin on it. Matthew Sadler and Natasha Regan, authors of the most complete book on AlphaZero's chess-playing, explained that it learned "in a unique manner by playing millions of lightning-fast games against itself. It was given no human knowledge about established chess strategy. As a result, AlphaZero was free to develop its own chess techniques and style" (Sadler & Regan, 2019, p. 434).[56] According to AI researcher, literary critic and chess player Manny Rainer, AlphaZero is more than a very strong chess engine, it is "a non-human agent who, on its own and in less than a day, has discovered some extremely deep and interesting things about a game that people have been playing for over a thousand years" (2019, ¶ 2). And just like Shannon and others envisaged and as the Drosophila metaphor indicates, it is clear that for Google's DeepMind, the team behind AlphaZero, chess is considered a gateway to bigger things:

> [I]t would be easy to forget that AlphaZero is about more than just chess. AlphaZero is a proof of concept, demonstrating AI's capacity to crack complex problems without the use of human knowledge of strategy. In other words, chess is the testing ground. (Sadler & Regan, 2019, p. 434)

Having been left in the dust by the new generation of game playing machines, what hope remains for the human species? Kasparov himself, once viewed as the champion upon whom the pride of humankind rested, points the way by urging us not to fall prey to pessimism, but rather to appreciate the possibilities for freedom and creativity that the close collaboration with our machines will open up: "I do not believe in fates beyond our control. Nothing is decided. None of us are spectators. The game is under way and we are all on the board" (Kasparov & Greengard, 2017, p. 136).

In their study of the future of work in the age of automation, Erik Brynjolfsson & Andrew McAfee, echoing Michie's earlier stated dream, concur: "the best chess player on the planet today is not a computer. Nor is it a human. The best chess player is a team of humans using computers" (Brynjolfsson & McAfee, 2011, p. 38). This leads us to the next section, where we present the optimistic view that our future relationships with machines may end up being cooperative rather than competitive or subservient.

Human-Machine Symbiosis

> We have let ourselves become enchanted by big data only because we exoticize technology. We're impressed with small feats accomplished by computers alone, but we ignore big achievements from complementarity because the human contribution makes them less uncanny. Watson, Deep Blue, and ever-better machine learning algorithms are cool. But the most valuable companies in the future won't ask what problems can be solved with computers alone. Instead, they'll ask: how can computers help humans solve hard problems? (Thiel & Masters, 2014, p. 83)

One of the proposals to deal with the risks of over-reliance in (if not complete annihilation at the hands of) machines is that of preemptively merging with them, so that as the capacities of standalone AIs increase, so do ours to keep them in check. The possibility is heralded in tech articles with headlines such as *Humans With Amplified Intelligence Could Be More Powerful Than AI,* which goes on to claim that:

> With much of our attention focused the rise of advanced artificial intelligence, few consider the potential for *radically amplified human intelligence* (IA). It's an open question as to which will come first, but a technologically boosted brain could be just as powerful—and just as dangerous—as AI. […] Unlike efforts to develop artificial general intelligence (AGI), or even an artificial superintelligence (SAI), *the human brain already presents us with a pre-existing intelligence to work with.* (Dvorsky, 2013, ¶ 1, emphases in the original)

The idea has gained traction recently, with, for instance, Elon Musk founding Neuralink a company aimed at improving brain-computer interfaces in order to let humans "achieve a sort of symbiosis with artificial intelligence" (Etherington, 2019, ¶ 2), but it has a long history. William Ross Ashby, creator of the Homeostat—referred to in the pages of Time Magazine as "the thinking machine" (Ramage & Shipp, 2009, p. 46) and "the closest thing to a synthetic brain so far designed by man" (Ashby, 2008, ¶ 30)—was hinting in this direction in his *Introduction to Cybernetics* when talking of the amplification of intellectual power, even if he did so in less than a page, before hurriedly ending the book:

> Now "problem solving" is largely, perhaps entirely, a matter of appropriate selection […] Thus it is not impossible that what is commonly referred to as "intellectual power" may be equivalent to "power of appropriate selection" […] If this is so, and as we know that power of selection can be amplified, it seems to follow that intellectual power, like physical power, can be amplified. Let no one say that it cannot be done, for the gene-patterns do it every time they form a brain that grows up to be something

> better than the gene-pattern could have specified in detail. What is new is that we can now do it synthetically, consciously, deliberately. But this book must stop; these are not matters for an Introduction. (Ashby, 1957, p. 272)

Three years later, J.C.R. Licklider published a paper in which he laid down the possibility for Man-Computer Symbiosis. But it is interesting to note that he expressed clear doubts as to whether in the long run these hybrid systems would be able to outperform a new generation of fully wetware-independent machines:

> Man-computer symbiosis is probably not the ultimate paradigm for complex technological systems. It seems entirely possible that, in due course, electronic or chemical "machines" will outdo the human brain in most of the functions we now consider exclusively within its province. (1960, p. 4)

But like Kasparov, many others have seen the enormous potential in the joint operation of man and machine. Frederick Brooks even goes as far as to say that this should have been the actual goal all along and that the quest for Artificial Intelligence was misdirected from the start in a way that set back the advance of computer science by sending researchers after a red herring:

> It is time to recognize that the original goals of AI were not merely extremely difficult, they were goals that, although glamorous and motivating, sent the discipline off in the wrong direction. If indeed our objective is to build computer systems that solve very challenging problems, my thesis is that […] *intelligence amplifying* systems can, at any given level of available systems technology, beat AI systems. That is, a machine *and* a mind can beat a mind-imitating machine working by itself. (1996, p. 64, emphases in the original)

Even Hubert Dreyfus, one of the harshest critics of Artificial Intelligence, seems not to have been against the scenario of man-machine integration and deems complementarity a more fruitful pursuit than automation, approvingly citing researchers advocating that "work be done on systems that promote a symbiosis between computers and human beings" (Dreyfus, 1965, p. 83): "Man and computer is capable of accomplishing things that neither of them can do alone" (Rosenblith, as cited in Dreyfus, 1965, p. 83).

Lyle Burkhead argues in the same vein that "whatever capability AI has at any given time, humans assisted by computers will have already reached that point and moved ahead" (Burkhead, 1999, p. 3). According to him, the reason for the advantages that human-machine teams would have over standalone machines consists in that the leap in machine intelligence that will take them from a human

to a superhuman level cannot be magical; machines will find the same epistemological ceiling that we have and will have to overcome it much like we would. There are no bootstrapping shortcuts and furthermore, whatever a machine operating by itself may be capable of accomplishing in this regard, a capable human aided by a machine will be able to do first, and better.

Such a possibility is no longer a speculation about the future. Many steps in that direction have already been taken and the road for super-intelligence via the aid of integrated machines is one we are already traversing: "External hardware and software supports now routinely give human beings effective cognitive abilities that in many respects far outstrip those of our biological brains" (Bostrom & Sandberg, 2009, p. 311). We are already living examples that the first preliminary steps that could lead in the direction of full merging have come to pass; the gadgets that we incorporate into our daily lives have endowed us with the possibility of doing things, communicating and accessing volumes of information in ways unthinkable to our ancestors:

> Computers are extensions of our minds, [they] are more than repositories for our memories and plans; they stand alone. They are half tool, half entity. […] Each new technology that humans adopt has the effect of amplifying our actions. Each new technology is a barrier removed between us and our ultimate freedom. As knives amplify and extend teeth and fingernails, as pliers amplify fingers, so do computers amplify our brains. Our identification with our computers marks the beginning of an incremental merging process whose end point will be a symbiosis of sorts. (Dewdney, 1998, p. 99)

This idea of enhancing ourselves, drastically altering who we are and how we interact with the world via the integration of semi-intelligent machines could just be one extreme expression of a basic feature that in fact defines the very essence of the relationship in which humans live amidst the world around them and those things that lie outside the barriers of their skin. Building on his and philosopher David Chalmers's previous idea of the extended mind (Clark & Chalmers, 1998), Andy Clark claims that we, as natural-born cyborgs, have an innate propensity to establish very close relationships with nonbiological resources and that the distinction between world and person is extremely—and increasingly more so—difficult to establish:

> The cyborg is a potent cultural icon of the late twentieth century. It conjures images of human-machine hybrids and the physical merging of flesh and electronic circuitry. My goal is to hijack that image and to reshape it, revealing it as a disguised vision of (oddly) our own biological nature. For what is special about human brains, and what best explains the distinctive features of human intelligence, is precisely

> their ability to enter into deep and complex relationships with nonbiological constructs, props, and aids. (Clark, 2003, p. 5)
>
> Human thought and reason is born out of looping interactions between material brains, material bodies, and complex cultural and technological environments. We create these supportive environments, but they create us too. We exist, as the thinking things we are, only thanks to a baffling dance of brains, bodies, and cultural and technological scaffolding. (Clark, 2003, p. 11)

No other gadget that we own shows this more vividly than our smartphones (which is a weird and outdated name for our portable personal pocket computers, as future historians will probably agree). Smartphones with machine intelligence aim to be "the part of your brain you're not born with" (Gershgorn, 2019, ¶ 59). This last statement is day by day seeming more literal than metaphorical, for smartphones highlight the tension in our relationship with technology, between complementarity and enrichment and dependency, which—after this long detour—brings us back into the province of games.

Gamified Us

A common misconception of folk paleontology is that all dinosaurs became extinct. While it is true that those huge lumbering beasts that are the delight of children all over the world no longer roam our plains nor traverse the watery realms of our oceans, dinosaurs surround us everywhere and not a day goes by without our meeting them. We simply call them 'birds'. A Tyrannosaurus rex is more akin to a chicken than to a Stegosaurus (other than in their sizes, of course, but most certainly when it comes to their morphology and the timespan separating their appearance on the evolutionary stage)[57] and anyone that has taken a minute to look a peacock in the eye, disregarding for a second the magnificent glimmer of its coat, will be able to attest that the old rulers of our planet linger still.

In a similar legend, behaviorists became officially extinct somewhere circa the late 1960s after Noam Chomsky arrived to save the day, dismounted from his generative horse and slew that foul beast, *Verbal Behavior*.[58] But the pigeons merrily strutting around should not only remind us of the stubborn subsistence of the scaly forebears they embody, but also that the insights of applied behaviorism, to which they so devotedly contributed, are all around us too: We encounter them first and foremost in the multibillion-dollar video game industry. Such a connection has been explicitly drawn out by Linehan, Kirman & Roche (2014, p. 82) in their chapter *Gamification as Behavioral Psychology*, where they explain how "the effects of characteristic game design elements (i.e., points, badges, leaderboards,

time constraints, clear goals, challenge) can be explained through principles of behavior investigated and understood by behavioral psychologists for decades (see Skinner 1974)."

With 2.5 billion people—that is, nearly a third of all human beings—playing video games (WePC, 2019) the industry is more lucrative than ever. Its earnings have for the past eight years far surpassed those of the movie and music industries combined (League of Professional eSports, 2018), and are expected to surpass 90 billion USD by 2020 (WePC, 2019). With that kind of money on the table, the slightest tweak that may help capture and retain players becomes invaluable, which is why companies are increasingly relying on psychologists in order to assist with game design. In the words of an article luring psychology grad students into the video game production world, "companies that design and develop video games are increasingly turning to psychologists for help analyzing data and making sure their products are as effective as they can be. Some psychologists are even launching consulting businesses to assist game manufacturers" (Clay, 2012, ¶ 2).

As it is with most things, there are shades of white and black in this relationship between psychology and video games, ranging from their immense potential to facilitate learning in an educational setting (Rosas et al. 2003) and the lofty goals of Game User Research "to improve player experience in games" (Nacke, 2018) to the more harrowing depths of employing behavioristic reward schedules in order to reinforce video game play so as to lead to an addictive relationship with them: "Like gambling on slot machines, video games reinforce correct or skilful play on variable and fixed ratio reinforcement schedules" (King, Delfabbro & Griffiths, 2009, p. 100). In an environment that is increasingly competitive, a strategy that game companies can greatly benefit from in order to remain profitable is that of exploiting the mental makeup and biases of their users, so as to make games irresistible (Søraker, 2016). With haunting vividness, creative writing teacher extraordinaire Jerome Stern describes the Sisyphean futility of being thus inescapably hooked:

> [M]y eyes stare intensely and my brain cells sizzle and fry. I am playing a computer game […] This hopelessly pointless game is slurping up thousands of life-seconds like a voracious anteater in a giant colony. My fingers dance on buttons and I can feel my time on earth being shortened, my vitality being sucked, my head spinning. I am using these fragile moments of our brief vanishing years, these precious minutes of lucidity that crumble sooner than we think, not to answer human correspondence, not to record my thoughts, not to do good in the world, but to press cd: GAME. GAME, and squawk goes the screen and little figures bounce out, pointlessly jump, and more moments of my life gasp like guppies and flop over gone and I can't help it. I can't stop. (Stern, 1997, p. 48)

However, much more worrisome is the fact that not only is the population at large far more reliant than ever on video games as a pastime, but that in an increasingly technological world, the boundaries between play and non-play grow ever more fluid, and life itself is becoming increasingly 'gamified', both explicitly and tacitly. As with any proprietary term of high potential profitability, there is ample contention as to what gamification entails precisely,[59] and while a nuanced survey of such disputes would be fruitful, it falls outside the scope of this essay. Yu Kai-Chou highlights its potential for good when defining it as "the craft of deriving fun and engaging elements found typically in games and thoughtfully applying them to real-world or productive activities" (2016, p. 8) while surveillance scholar Jennifer Whitson calls attention to a more troublesome aspect of the practice, for in her definition gamification "applies playful frames to non-play spaces, leveraging surveillance to evoke behaviour change" (2013, p. 164). In an astute simile, Jason Fagone adds that "gamification advocates—like religious figures—seek to superimpose an invisible reward system on top of the world" (Fagone, 2011, ¶ 6). The harshest criticism to gamification comes from game designer Ian Bogost, who channeling Harry Frankfurt's poignant rhetoric and illuminating analysis from *On Bullshit*, accuses its proponents of being bullshit peddlers trying to lure business executives with a sexy buzzword that ends up being nothing but a front for very old practices (Bogost, 2014). Having shared these words of caution and bearing such valid concerns in mind, we acknowledge that we use the term in a looser and more encompassing way than is common.

In a striking example of the trend, Amazon recently rolled out a video game-like interface that reflects the progress that warehouse workers are making at their tasks, while other companies like Uber, Delta Air Lines and Target have employed gamification in turn to affect their own metrics (Bensinger, 2019). Given reports that have emerged of the poor working conditions at Amazon warehouses, the retail giant would seem like a prime source of validity for Ian Bogost's critique that proposes substituting the term 'gamification' with 'exploitationware', owing to its replacing "real, functional, two-way relationships with dysfunctional perversions of relationships. Organizations ask for loyalty, but they reciprocate that loyalty with shams, counterfeit incentives that neither provide value nor require investment" (2011, ¶ 57). But why should companies stop at their employees, when there is so much profit to be reaped by gamifying consumers too? Loyalty programs can be seen as a form of proto-gamification, and with the ongoing sophistication of technology, we should expect them to become increasingly pervasive, with, for instance, Netflix or Amazon framing certain landmarks in book-buying or episode-watching as epic quests, appropriately rewarded by a badge or some other such sign.

And that gamified nature may be already embedded in our relationship with the technological tools we employ the most:

> Jamie Madigan, a psychologist who writes about video games, thinks the arrival of a notification might be similar to the accrual of virtual loot. Email, in other words, might not be just a task, but a game. "Designers of apps for the Web, phones, and other devices figured this out early on," he says. "In the case of our phones, we see, hear, or feel a notification alert show up, we open the app, and we are rewarded with something we like: a message from a friend, a like, an upvote, or whatever." (Pinsker, 2015, ¶ 8)

And this reinforcing quality of the technology we use daily is certainly a feature, not a bug. As Will Chamberlain puts it in discussing recently proposed legislation to tackle the issue, "the problem isn't just that social media use can be addictive; the problem is that it's designed to be addictive" (Chamberlain, 2019, ¶ 10). Gamified aspects of ubiquitous technology can be even more subtle. "Not a few futurologists envisage a network of computer users tired, apparently, of violent or 'erototronic' video games engaging, instead, in political debate; a hi-tech resurrection, on a grand scale, of the participatory democracy of the Athenian agora" claimed philosopher David E. Cooper (1995, p. 10) presciently prefiguring Twitter eleven years before its creation. The metrics (retweets, follower count, etc.) have a gamified flavor that warrants reading the platform as a political video game of sorts. And by the same token, a dating app such as Tinder can also be better understood as a video game, where for many users the 'match' is an end in itself as an ego-boost, regardless of whether any subsequent meeting up in physical space actually occurs.

But while the Skinnerian conditioning that we receive from our devices may fly completely under the radar for some of us, others pursue it of their own accord. That is what we see in the case of Piotr Wozniak, developer of the memory-aiding program SuperMemo, by which he rules his life (Wolf, 2008). The program stores every bit of information and every new fact that Wozniak judges important or worth preserving, and, in a manner fully reminiscent of Ebbinghaus' theories of memory, then presents it again and again at precisely spaced intervals, until they have been completely learned. Wozniak takes his reliance on the program to an extreme degree and turns over the administration of his life to his personally designed computer system. Such decisions as what to read, what to re-read and when, who to see, who to reply to, are routinely decided by the software that he has devoted most of his life to develop and perfect:

> When he entrusts his mental life to a machine, it is not to throw off the burden of thought but to make his mind more swift. Extreme knowledge is not something for which he programs a computer but for which his computer is programming him. (Wolf, 2008, p. 10)

Wozniak's example is particularly striking for the usual bonds between creator and creature are thrown into a loop, with the programmer making his machine and then being re-made by it. But however extreme the case of Wozniak may seem to us, the fact is that the infrastructure is set in place for such kinds of relationships between humans and programs to be far more common. Here is Yuval Noah Harari, with a forecast worth considering, if not for the forecaster's sapience, at the very least owing to the widespread attention and success with which the book in which it appears, *Homo Deus: A Brief History of Tomorrow*, has been met:

> Companies such as Mindojo are developing interactive algorithms that not only teach me maths, physics and history, but also simultaneously study me and get to know exactly who I am. Digital teachers will closely monitor every answer I give, and how long it took me to give it. Over time, they will discern my unique weaknesses as well as my strengths. (2016, p. 163)

This harbinger sign of our willingness to give in the reins of our mental development to algorithms brings to mind Martin Heidegger's words of caution to the effect that what is at stake in our dealings with machines is not so much our worldly hegemony but ourselves: "The threat to man does not come in the first instance from the potentially lethal machines and apparatus of technology. The actual threat has already affected man in his essence" (Heidegger, 1977, p. 28). Of course, stressing the importance of Heidegger's admonition should not in the least lead us to disregard the true threat represented by the "potentially lethal machines and apparatus of technology" (see [MYTHS] for a fuller treatment of such risks) but rather to not lose sight that, as pointed out by concerned contemporary writers on technology, "as we come to rely on computers to mediate our understanding of the world, it is our own intelligence that flattens into artificial intelligence" (Carr, 2008, ¶ 37).

But if there is one arena in which the risks to humankind's survival, the risks for the survival of its humanity, and the feedback loops between games and technology all come into play is in modern military warfare. "War games have been serious business for military leaders over the years," declares media researcher and game developer Casey O'Donnell (2014, p. 351). The very intimate relationship between video games and armed conflict is well documented (see Mead, 2013) and expresses itself in several ways. A noteworthy example is *America's Army*, a first-person shooter developed by the US

Army that attempts to portray combat situations more realistically than other franchises, and is intended mainly as a recruitment tool (Allen, 2014). But *America's Army* is far from the only video game that soldiers will be playing: "United States troops stationed overseas […] dedicate so many hours a week to burnishing their Halo 3 in-game service record that earning virtual combat medals is widely known as the most popular activity for off-duty soldiers" (McGonigal, 2010, p. 8). Nicole Capezza, extending important work by Jaan Valsiner, draws a crucial implication of the video game-like ethos of contemporary warfare from a cultural psychology standpoint by means of the concept of distancing:

> During wartime soldiers often use distancing mechanisms when deciding whether or not to shoot at an "enemy" soldier. New mechanisms for psychological distancing are making these decisions easier. Night-vision or thermal imagery converts the "enemy" soldier into, "an inhuman green blob." This technology and the distancing process have been referred to as "Nintendo warfare." (Capezza, 2003, ¶ 22)

The fact that much killing can now be conducted with an added layer of detachment (i.e., via piloting drones remotely) makes this an even more worrisome reality. There is nevertheless some bitter consolation to be had in the fact that trends point to an increasing automation of lethal weapons, with a push for drones being able to employ lethal force without human oversight. This is such a concerning possibility that many of the world's leading AI researchers and technologists have signed an open letter urging authorities not to start an AI arms race by the creation and deployment of autonomous weapons (Future of Life Institute, 2015).

Endgame

It is now time to return to Artificial Intelligence to tie up what we have been discussing with our initially proffered suspicions that the gamified nature of our technological milieu may eventually usher in a future in which, similarly to the case of Piotr Wozniak, it is our machines who create the games in which we are subsumed, a future consisting of an infantilized humankind being watched over by AIs.

Jane McGonigal, a cheerful, thoughtful and well-meaning evangelist for the positive power of gameful design, begins her largely optimistic account of the social future and transformative potential of video games, *Reality Is Broken: Why Games Make Us Better and How They Can Change the World,* with this passage from Edward Castronova's *Exodus to the Virtual World*:

> Anyone who sees a hurricane coming should warn others. I see a hurricane coming. Over the next generation or two, ever larger numbers of people, hundreds of millions, will become immersed in virtual worlds and online games. While we are playing, things we used to do on the outside, in "reality," won't be happening anymore, or won't be happening in the same way. You can't pull millions of person-hours out of a society without creating an atmospheric-level event. (cited in McGonigal, 2010, p. 8)

We hope the foregoing discussion has lent credence to this possibility and we expect the continuous reporting on the improvement of virtual reality technology to do as much. A crucial question, however, is: Whose virtual worlds? Whose online games? Earlier, we cited cybernetician W. Grey Walter on the species-wide developmental impact of play, but the follow-up to his comment on the importance of games and play for our civilization is equally worth taking into account, if not more so: "Perhaps the most ominous feature of mechanized civilization is that the ludicrous devices demanded for entertainment do not lend themselves to two-way operation" (Walter, 1961, p. 225). That is, the means of distraction are handed top-down, and ultimately, consumers have very little input in their design. It is at the hands of the algorithms initially rolled out by the State and corporations, but then liable of cutting such ties with their creators (see [MYTHS]).

Should we cruise along our current path, we may soon be facing a similar scenario to that of the Merovingian *rois fainéants*, or do-nothing kings, who gave up to their Carolingian majors of the palace the administration of their affairs, with little protest, in the pursuit of their own forms of entertainment. Being a century and a half ahead of his time, Samuel Butler already outlined how such a gradual shift might unfold, in a manner reminiscent of the fabulaic demise of the slowly-boiling frog of lore:

> The power of custom is enormous, and so gradual will be the change, that man's sense of what is due to himself will be at no time rudely shocked; our bondage will steal upon us noiselessly and by imperceptible approaches: nor will there ever be such a clashing of desires between man and the machines as will lead to an encounter between them.[60] (Butler, 1872/2014, p. 81)

The looming threat having now drawn considerably closer, we hear echoes of that very concern that the shift will occur in steps so gradual as to be functionally imperceptible until it is truly too late in the writings of Theodore Kaczynski, who sought to force attention to be paid to the threats he perceived in the raising technologization of society by means both textual and paratextual. Here is part of proposition 173 in his manifesto, published under coercion in 1995 by *The Washington Post*:

> What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and as machines become more and more intelligent, people will let machines make more and more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide. (Kaczynski, 2010, p. 77)[61]

That we would be able to hand command of our lives over to machines should not be altogether too surprising given that we already have in place precisely such a blueprint of dependence, although, to a different owner. We appreciate uncanny resemblances between the kind of technological world we have been describing as partially in existence, and a flight of sci-fi full of foresight that Alexis de Tocqueville encoded in his 1840 fourth volume of *Democracy in America*, which philosopher Anthony O'Hear (1995, p. 158) credits as "the most accurate portrait of our age":

> Above this race of man stands an immense and tutelary power, which takes it upon itself alone to secure their gratifications and to watch over their fate. That power is absolute, minute regular provident and mild. It would be like the authority of a parent if, like that authority, its object was to prepare men for manhood; but it seeks on the contrary, to keep them in perpetual childhood: it is well content that people should rejoice provided that they think of nothing but rejoicing… (in O'Hear, 1995, p. 158)

O'Hear then goes on to make the parallels between de Tocqueville's anticipation and our current technologically-infused world all the more explicit, by saying that "technology infantilizes, encouraging people to be satisfied with the material delights it makes so easy, and to reduce our sense of freedom and democracy to that of chosing among the delights and 'life-styles' they make possible" (1995, p. 158). But frankly, and as valuable as his analysis truly is, he needn't even have bothered, for the similarities between what de Tocqueville foresaw and the gamified ecosystems we have been talking about are so on-the-nose that stressing them would seem to be no more than belaboring the point:

> So I think that the type of oppression by which democratic peoples are threatened will resemble nothing of what preceded it in the world; our contemporaries cannot find the image of it in their memories. I seek in vain myself for an expression that exactly reproduces the idea that I am forming

> of it and includes it; [<the thing that I want to speak about is new, and men have not yet created the expression which must portray it.>] the old words of despotism and of tyranny do not work. The thing is new, so I must try to define it, since I cannot name it. I want to imagine under what new features despotism could present itself to the world; I see an innumerable crowd of similar and equal men who spin around restlessly, in order to gain small and vulgar pleasures with which they fill their souls. Each one of them, withdrawn apart, is like a stranger to the destiny of all the others; his children and his particular friends form for him the entire human species; as for the remainder of his fellow citizens, he is next to them, but he does not see them; he touches them without feeling them; he exists only in himself and for himself alone. (de Tocqueville, 2010, p. 1250)[62]

Verily, those "small and vulgar pleasures" sound eerily reminiscent to the empty badges, points and achievements that the critics of gamification justly decry. In autodidact sociologist Eli Sagan's comparison between ancient Greek and modern American democracies, there is a passage of great explanatory power that seems to be looking deeper into our collective psychological relationship to that 'tutelary power' so vividly depicted by de Tocqueville:

> The collectivized person is also constantly struggling with the universal human ambivalence about independence and dependence. Like a child, the *demos* longs to put its entire trust in the hands of its leaders, becoming enraged when the leaders, like parents, fail to deliver omnipotence, omniscience, or moral perfection. This disenchantment does not prevent the pattern from being repeated over and over again. […] The desire to be illusioned runs very deep in the human psyche. (Sagan, 1991, p. 195)

Indeed, so strong is our desire for illusions that we may end up permanently confined to comforting illusions (comforting, that is, when confronted with the dismal alternative of a bleak and threatening reality), as very popular films such as *The Matrix* and *The Truman Show* (or Stanisław Lem's novel *The Futurological Congress* (1971), the masterpiece of the simulacra genre) have portrayed and which have so gripped the imagination and influenced contemporary discussion that they have forced academic philosophers to take them up as serious objects of concern (Chalmers, 2005). And, as Huizinga himself explained, 'illusion' is "a pregnant word which means literally 'in-play' (from *inlusio*, *illudere* or *inludere*)" (1980, p. 11).

References

Aldunate Phillips, A. (1964). *Los robots no tienen a Dios en el corazón*. Santiago, Chile: Editorial Andrés Bello.

Allen, R. (2014). America's Army and the Recruitment and Management of 'Talent': An Interview with Colonel Casey Wardynski. *Journal of Gaming & Virtual Worlds*, Volume 6, Number 2.

Ashby, W.R. (1957). *An Introduction to Cybernetics*. London: Chapman & Hall Ltd.

Ashby, J. (2008). *Biography: W. Ross Ashby (1903-1972)*. Retrieved from: http://www.rossashby.info/biography.html

Augustine, S. (1876). *The Confessions of S. Augustine*. [Revised from a former translation by E.B. Pusey]. London: James Parker & Co.

Babbage, C. (1864). *Passages from the Life of a Philosopher*. London: Longman, Green.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B. & Mordatch, I. (2019). Emergent Tool Use From Multi-Agent Autocurricula. *ArXiv e-prints*. arXiv:1909.07528

Bensinger, G. (2019, May 21). 'MissionRacer': How Amazon turned the tedium of warehouse work into a game. *The Washington Post*.

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford: Clarendon Press.

Bogost, I. (2011). Persuasive Games: Exploitationware. Gamasutra, May 3. Retrieved from: https://www.gamasutra.com/view/feature/134735/persuasive_games_exploitationware.php

Bogost, I. (2014). Why Gamification is Bullshit. In S.P. Walz & S. Deterding (Eds.) *The Gameful World: Approaches, Issues, Applications*. Cambridge, MA: MIT Press, pp 65–79.

Bruns, A. & Jacobs, J. (2007). *Uses of Blogs*. NY: Peter Lang.

Brynjolfsson, E. & McAfee, A. (2011). *Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Lexington, MA: Digital Frontier Press.

Burghardt, G. (2005). *The Genesis of Animal Play: Testing the Limits*. Cambridge, MA: The MIT Press.

Burkhead, L. (1999). The Irrelevance of AI -- the Reverse Turing Test. In L. Burkhead (Auth.) *Nanotechnology Without Genies*. Retrieved from: http://www.geniebusters.org/21_Turingtest.htm

Butler, S. (2014). The Book of the Machines. In R. Mackay & A. Avanessian (Eds.) *#ACCELERATE#: The Accelerationist Reader*. Falmouth, UK: Urbanomic, pp. 67–90. (Original work published 1872).

Caillois, R. (2001). *Man, Play and Games*. Urbana and Chicago: University of Illinois Press.

Capezza, N. (2003). The Cultural-Psychological Foundations for Violence and Nonviolence. An Empirical Study. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 4(2). doi:http://dx.doi.org/10.17169/fqs-4.2.717

Carr, N. (2008). Is Google Making Us Stupid? What the Internet is doing to our brains. *The Atlantic*. July/August 2008 Issue. Retrieved from: https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/

Carr, N. (2015). The Control Crisis. In J. Brockman (Ed.), *What to Think About Machines That Think*. NY: HarperCollins. pp. 59-61.

Carse, J.P. (1986). *Finite and Infinite Games*. NY: Free Press.

Chalmers, D 2005, The Matrix as Metaphysics. In C. Grau (Ed.), *Philosophers Explore The Matrix*. NY: Oxford University Press, pp. 132-176.

Chamberlain, W. (2019, August 26). Josh Hawley Is Right About Social Media Addiction. *Human Events*. Retrieved from: https://humanevents.com/2019/08/26/josh-hawley-is-right-about-social-media-addiction

Chou, Y. (2016). *Actionable Gamification: Beyond Points, Badges and Leaderboards*. Fremont, CA: Octalysis Media.

Christian, B. (2011). *The Most Human Human: What Artificial Intelligence Teaches Us About What it Means to Be Alive*. NY: Doubleday.

Clay, R.A. (2012). Video game design and development. *GradPSYCH Magazine*, January 2012.

Cole, N. (2016). Confessions of a Teenage Gamer. *Nicolas Cole*.

Cooper, D. (1995) Technology: Liberation or Enslavement? In R. Fellows (Ed.) *Philosophy and Technology*. Cambridge: Cambridge University Press, pp. 7-18.

Copeland, B.J. (2004). *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press.

Copeland, B.J. (2017). Computer chess—the first moments. In B.J. Copeland, J. Bowen, M. Sprevak, & R. Wilson (Eds.) *The Turing Guide*. Oxford: Oxford University Press, pp. 327–346.

Contreras, J.I. (2019) Playing with pattern: Aesthetic communication as distributed cognition. *Aisthesis* 12(1): 27-39.

Conway, F. & Siegelman, J. (2005). *Dark Hero of the Information Age: In Search f Norbert Wiener, the Father of Cybernetics*. NY: Basic Books.

Darwin, C.G. (1952). *The Next Million Years*. London: Rupert-Hart Davies.

de Tocqueville, A. (2010). *Democracy in America: Historical-Critical Edition of 'De la démocratie en Amérique'*. (Edited by Eduardo Nolla and translated by James T. Schleifer). Indianapolis: Liberty Fund.

Dreyfus, H. L. (1965). *Alchemy and AI*. [Technical report]. Santa Monica, CA: RAND Corporation.

du Sautoy, M. (2019). *The Creativity Code: How AI is Learning to Write, Paint and Think*. London: 4th Estate.

Dvorsky, G. (2013). Humans With Amplified Intelligence Could Be More Powerful Than AI. *io9*. Retrieved from http://io9.com/humans-with-amplified-intelligence-could-be-more-powerf-509309984

Ekbia, H. (2008). *Artificial Dreams: The Quest for Non-Biological Intelligence*. Cambridge: Cambridge University Press.

Ekbia, H. & Nardi, B. (2017). *Heteromation, and Other Stories of Computing and Capitalism*. Cambridge, MA: MIT Press.

Ensmenger, N. (2011). Is chess the drosophila of artificial intelligence? A social history of an algorithm. *Social Studies of Science*, 42(1), 5–30.

Etherington, D. (2019, July 16). Elon Musk's Neuralink looks to begin outfitting human brains with faster input and output starting next year. *TechCrunch*. Retrieved from: https://techcrunch.com/2019/07/16/elon-musks-neuralink-looks-to-begin-outfitting-human-brains-with-faster-input-and-output-starting-next-year/

Evans, H.R. (1906). The Romance of Automata. In H.R. Evans (Auth.) *The Old and the New Magic*. Chicago: The Open Court Publishing Co, pp. 131–141.

Fagone, J. (2011, July 15). Chain World Videogame Was Supposed to be a Religion—Not a Holy War. *WIRED*. Retrieved from: https://www.wired.com/2011/07/mf_chainworld/

Future of Life Institute. (2015, July 28). *Autonomous Weapons: An Open Letter From Ai & Robotics Researchers*. Retrieved from: https://futureoflife.org/open-letter-autonomous-weapons/

Gershgorn, D. (2019, April 16). How Google Aims To Dominate Artificial Intelligence. *Popular Science*. Retrieved from: https://www.popsci.com/google-ai/

Gilbert, D. (2006). *Stumbling on Happiness*. NY: Alfred A. Knopf.

Grimstad, P. (2013). *Experience and Experimental Writing: Literary Pragmatism from Emerson to the Jameses*. Oxford: Oxford University Press.

Hanson, R. (2016). *The Age of Em: Work, Love and Life when Robots Rule the Earth*. Oxford: Oxford University Press.

Harari, Y.N. (2016). *Homo Deus: A Brief History of Tomorrow*. UK: Harville Secker.

Harari, Y.N. (2018). *21 Lessons for the 21st Century*. NY: Spiegel & Grau

Heidegger, M. (1977). *The Question Concerning Technology and Other Essays*. [Translated by William Lovitt]. NY: Garland Publishing, Inc.

Hofstadter, D. (1995). Retrieval of Old and Invention of New Analogies. In D. Hofstadter and the Fluid Analogies Research Group (Eds.) *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. NY: Basic Books.

Hofstadter, D. (2000). *Gödel, Escher, Bach: An Eternal Golden Braid*. London: Penguin. (Original work published 1979).

Holland, O. (2003). The first biologically inspired robots. *Robotica*, volume 21, pp. 351–363.

Huizinga, J. (1980). *Homo Ludens: A Study of the Play Element in Culture*. London: Routledge.

Hurtado, E. (2017). *Consequences of Theoretically Modeling the Mind as a Computer*. (Doctoral dissertation). Pontificia Universidad Católica de Chile.

Jacobellis v. Ohio, 378 U.S. 184 (1964)

James, W. (1950). *The Principles of Psychology: Volume One*. NY: Dover. (Original work published 1890).

Kaczynski, T. (2010). Industrial Society and Its Future. In T. Kaczynski & D. Skrbina (Eds.) *Technological Slavery: the Collected Writings of Theodore J. Kaczynski, a.k.a. 'The Unabomber'*. NY: Feral House, pp. 27–151.

Kamb, S. (2016). *Level Up Your Life: How to Unlock Adventure and Happiness by Becoming the Hero of Your Own Story*. Pennsylvania: Rodale Books.

Kasparov, G. & Greengard, M. (2017). *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. NY: PublicAffairs.

King, D., Delfabbro, P., & Griffiths, M. (2009). Video Game Structural Characteristics: A New Psychological Taxonomy. *International Journal of Mental Health and Addiction*, 8(1), 90–106.

Kline, R. (2011). Cybernetics, automata studies, and the Dartmouth conference on artificial intelligence. *IEEE Annals of the History of Computing*, 33(4), 5-16.

Kline, R. (2015). *The Cybernetics Moment: Or Why We Call Our Age the Information Age*. Baltimore: Johns Hopkins University Press.

League of Professional eSports. (2018, October 10). The Video Games' Industry is Bigger Than Hollywood. *LPEsports.com*. Retrieved from: https://lpesports.com/e-sports-news/the-video-games-industry-is-bigger-than-hollywood

Lem, S. (1974). *The Futurological Congress*. [Translated by Michael Kandel]. NY: Seabird Press.

Levi, D. & Newborn, M. (2012). *All About Chess and Computers*. Berlin: Springer Verlag.

Madrigal, A. (2017, July 19). How Checkers Was Solved. *The Atlantic*. Retrieved from: https://www.theatlantic.com/technology/archive/2017/07/marion-tinsley-checkers/534111/

McCorduck, P. (1979). *Machines Who Think*. San Francisco: W. H. Freeman and Company.

McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

McGonigal, J. (2010). *Reality is Broken: Why Games Make Us Better and How They Can Change the World*. NY: Penguin.

McGonigal, J. (2014). I'm Not Playful, I'm Gameful. In S.P. Walz & S. Deterding (Eds.) *The Gameful World: Approaches, Issues, Applications*. Cambridge, MA: MIT Press, pp 653–657.

McGonigal, J. (2015). *SuperBetter: A Revolutionary Approach to Getting Stronger, Happier, Braver and More Resilient--Powered by the Science of Games*. NY: Penguin.

Mead C. (2013), *War Play: Videogames and the Future of Armed Conflict*. Boston, MA: Houghton Mifflin Harcourt.

Michie, D. (1972). Programmer's gambit. *New Scientist*, 17 August 1972.

Minsky, M. (1984). Afterword to 'True Names'. In V. Vinge *True Names*. NY: Bluejay Books.

Munroe, R. (2013, May). Birds and Dinosaurs. *XKCD Comic*. Retrieved from: https://xkcd.com/1211

Nacke, L.E. (2018). Introduction to Biometric Measures for Games User Research. In A. Drachen, P. Mirza-Babaei & L.E. Nacke (Eds.) *Games User Research*. Oxford: Oxford University Press, pp. 281–299.

O'Donnell, C. 2014. Getting Played: Gamification, Bullshit, and the Rise of Algorithmic Surveillance. *Surveillance & Society* 12(3): 349-359.

O'Hear, A. (1995). Art and Technology: An Old Tension. In R. Fellows (Ed.) *Philosophy and Technology*. Cambridge: Cambridge University Press, pp. 143–158.

Organ, C. L., Schweitzer, M. H., Zheng, W., Freimark, L. M., Cantley, L. C., & Asara, J. M. (2008). *Molecular Phylogenetics of Mastodon and Tyrannosaurus rex*. Science, 320(5875), 499–499. doi:10.1126/science.1154284

Papert, S. (1968). *The Artificial Intelligence of Hubert L. Dreyfus: A Budget of Fallacies*. MIT AI Lab Memo 154. Cambridge, Mass.: MIT

Pasquinelli, M. (2017). Machines that Morph Logic: Neural Networks and the Distorted Automation of Intelligence as Statistical Inference. *Glass Bead*, 1: "Logic Gate: The Politics of the Artifactual Mind". Retrieved from: www.glass-bead.org/article/960

Pfannebecker, M. (2012). Cyborg Coriolanus/ Monster Body Politic. In S. Herbrechter & I. Callus (Eds.) *Posthumanist Shakespeares*. Basingstoke: Palgrave Macmillan Ltd., pp. 114-132.

Pinsker, J. (2015, May 27). Inbox Zero vs. Inbox 5,000: A Unified Theory. *The Atlantic*. Retrieved from: https://www.theatlantic.com/technology/archive/2015/05/why-some-people-cant-stand-having-unread-emails/394031/

Purves, D. (2019). Opinion: What does AI's success playing complex board games tell brain scientists? *Proceedings of the National Academy of Sciences*, 116(30), 14785–14787.

Ramage, M., & Shipp, K. (2009). *Systems Thinkers*. London: Springer.

Rayner, M. (2019). Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI. [Review]. *Goodreads*. Retrieved from: https://www.goodreads.com/review/show/2731237101

Ringeling, X. (2011). *El valor cognoscitivo de la metáfora: Su significado, su contenido, su verdad.* (Dissertation). Pontificia Universidad Católica de Chile.

Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., … Salinas, M. (2003). Beyond Nintendo: design and assessment of educational video games for first and second grade students. *Computers & Education*, 40(1), 71–94.

Sagan, E. (1991). *The Honey and the Hemlock: democracy and paranoia in ancient Athens and modern America*. NY: Basic Books.

Sadler, M & Regan, N. (2019). *Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI*. Alkmarr, The Netherlands: New in Chess.

Shannon, C. (1950). Programming a computer for playing chess. *Philosophical Magazine*, Ser.7, Vol. 41, No. 314.

Shannon, C. (1956). A Chess-Playing Machine. In J.R. Newman (Ed.) *The World of Mathematics*. NY: Simon & Schuster, p. 2124-2133.

Skinner, B.F. (1971). A Lecture on "Having" a Poem. In B.F. Skinner (Ed.) *Cumulative Record: A Selection of Papers*. (Third Edition). NY: Appleton-Century-Crofts, pp. 345–358.

Skinner, R. (2016). Artificial Intelligence. In H. Lowood & R. Guins (Eds.) *Debugging Game History: A Critical Lexicon*. Cambridge, MA: MIT Press, pp. 29-36.

Skrbina, D. (2010). A Revolutionary for Our Times. In T. Kaczynski & D. Skrbina (Eds.) *Technological Slavery: the Collected Writings of Theodore J. Kaczynski, a.k.a. 'The Unabomber'*. NY: Feral House, pp. 14–26.

Søraker, J. H. (2016). Gaming the gamer? – The ethics of exploiting psychological research in video games. *Journal of Information, Communication and Ethics in Society*, 14(2), pp. 106–123.

Standage, T. (2002a). *The Turk: The Life and Times of the Famous Eighteenth-Century Chessplaying Machine*. NY: Walker & Company.

Standage, T. (2002b, March 1). Monster in a Box. *WIRED*. Retrieved from: https://www.wired.com/2002/03/turk/

Stern, J. (1997). Game. In J. Stern (Auth.) *Radios: Short Takes on Life and Culture*. NY: W.W. Norton & Company, pp. 48–49.

Szymborska, W. (1981). *Sounds, Feelings, Thoughts: Seventy Poems by Wisława Szymborska*. [Translated and Introduced by Magnus J. Krynski and Robert A. Maguire]. New Jersey: Princeton University Press.

Tesler, L. (2019). *Larry Tesler CV: Adages & Coinages*. Retrieved from: http://www.nomodes.com/Larry_Tesler_Consulting/Adages_and_Coinages.html

Thiel, P. & Masters, B. (2014). *Zero to One: Notes on startups, or how to build the future.* New York: Crown Business.

Todd, A. (2016). Why Gamification is ~~Bullshit~~ Malarkey. *The Morning Watch: Educational and Social Analysis*. Vol 44, No 1-2 Fall.

Turing, A.M. (1948). Intelligent Machinery. In B.J. Copeland (Ed.) *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press, pp. 410–432.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind*, 59 (236), pp. 433–460.

Walter, W.G. (1961). *The Living Brain*. UK: Penguin Books.

WePC. (2019, June). 2019 Video Game Industry Statistics, Trends & Data. *WePC.com*. Retrieved from: https://www.wepc.com/news/video-game-statistics/#video-gaming-industry-overview

Whitson, J. (2013). Gaming the Quantified Self. *Surveillance & Society* 11 (1/2): 163-176.

Wiener, N. (1964) *God & Golem, Inc.: A Comment on Certain Points Where Cybernetics Impinges on Religion.* Cambridge, MA: MIT Press.

Wolf, G. (2008). Want to Remember Everything You'll Ever Learn? Surrender to This Algorithm. *WIRED*. Retrieved from: http://www.wired.com/medtech/health/magazine/16-05/ff_wozniak

Wolfram, S. (2017). A century of Turing. In B.J. Copeland, J. Bowen, M. Sprevak, & R. Wilson (Eds.) *The Turing Guide*. Oxford: Oxford University Press, pp. 43–47.

Ziff, P. (1959). The Feelings of Robots. *Analysis*, 19(3), 64–68.

Zyda, M. (2016). Why the VR You See Now Is Not the Real VR. *Presence*, Vol. 25, No. 2, Spring 2016, 166–168.

General Discussion: Creators, Creatures and Creativity

Roberto Musa G.

"[T]he isolated man does not develop any intellectual power. […]
From this point of view the search for new techniques
must be regarded as carried out
by the human community as a whole,
rather than by individuals."

Alan Turing, 1948, p. 431

Looking back on the three foregoing essays, there is a common thread running through them which I feel is worth bringing more explicitly into the foreground. It relates to the question of whether it is warranted to ascribe creativity to the productions and behaviors of machines. I note that in an earlier version of this very question I had phrased it as "whether the output of machines can be considered creative". I choose to point this out in order to draw attention to how subtle cues can already prejudge the answers we're liable to find (as was discussed in [TESTS]). 'Output' is a word with undeniably cold and mechanical undertones, and calling something an 'output' tacitly induces the reader to accept *a priori* that it has been made unthinkingly.

Throughout this discussion, I shall attempt to recast some of the main tenets of the three essays in this light, so that we can explore what conclusions they have to offer taken all together. We began our inquiry with [MYTHS], where we saw that a fundamental driving force behind the AI enterprise was that of emulating the mythical role of a divine creator. Attaining the power of creating autonomous minds seems to be one of the main allures that move AI researchers in their quest, and many of them are quite open about this influence. We have tried to offer as much evidence as we could muster for this assertion, but let us once more hark back to Norbert Wiener's words on the matter, when he explores at length the analogies between God and His creation, on the one hand, and Man and the machines he plays games with, on the other:

> The subject […] of machines that learn to play games […] is the problem of the game between the Creator and a creature (1964, p. 15) […] Can God play a significant game with his own creature? Can any creator, even a limited one, play a significant game with his own creature? In constructing machines

> with which he plays games, the inventor has arrogated to himself the function of a limited creator, whatever the nature of the game-playing device that he has constructed. (1964, p. 17)

It is significant that Wiener should ask whether a game played against one of our creatures can ever be 'significant', as this has been a bone of contention that accompanies AI ever since its prehistory. Machine learning pioneer Arthur Samuel created one of the first and most historically important game playing programs, his Checkers Player, but he had views completely antithetical to Wiener's on the question of the originality of the machine's play:

> A machine is not a genie, it does not work by magic, it does not possess a will, and, Wiener to the contrary, nothing comes out which has not been put in […] The "intentions" which the machine seems to manifest are the intentions of the human programmer, as specified in advance, or they are subsidiary intentions derived from these, following rules specified by the programmer. […] To believe otherwise is either to believe in magic or to believe that the existence of man's will is an illusion and that man's actions are as mechanical as the machine's. Perhaps Wiener's article and my rebuttal have both been mechanically determined, but this I refuse to believe. (Samuel cited in Hofstadter 1979/2000, p. 684)

In talking about his own chess playing program, Claude Shannon too admits that his machine "makes decisions, but the decisions were envisaged and provided for at the time of design. In short, the machine does not, in any real sense, go beyond what was built into it" (Shannon, 1956, p. 2133). Arturo Aldunate Phillips affirms essentially the same thing when he states that every move the chess program may make is "potentially known" (1964, p. 137) by whomever invented the mechanism. This is a very well known argument referred to as Lady Lovelace's objection, after the delightfully quirky Ada Byron, Countess of Lovelace, who first set it in stone (so to speak). Alan Turing tackled her objection head on, so let us see how he defends against it. He begins by quoting the Countess, who said: "The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform" (cited in Turing, 1950, p. 450, emphases hers). He then moves on to consider an alternative formulation of the objection:

> A variant of Lady Lovelace's objection states that a machine can "never do anything really new." This may be parried for a moment with the saw, "There is nothing new under the sun." Who can be certain that "original work" that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles. (1950, p. 450)

Let us pause here for a second to mention that far from being easy to disregard as an old saw, the wisdom from the Ecclesiastes of there being "nothing new under the sun" imbues the discussion spanning the three previous essays and shall land a center role in our reflection on the possibilities of machine creativity. After all, if we want to inquire as to whether machines can be creative, we need a standard against which to compare them. And let us not forget that even a thinker like Wittgenstein, also turning to the horticultural metaphor, was not all that confident of the extent of his own originality:

> My originality (if that is the right word) is, I believe, an originality that belongs to the soil, not the seed. (Perhaps I have no seed of my own.) Sow a seed in my soil, & it will grow differently than it would in any other soil. (Wittgenstein, 1998, p. 42)

Countess Lovelace's objection has led mathematician Marcus du Sautoy to propose, in the opening chapter of his reflections on machines, art and creation, *The Creativity Code*, that the Turing Test be complemented by a Lovelace Test as a benchmark for thinking. In order to pass it, an algorithm should be able to "originate a creative work of art such that the process is repeatable (i.e. it isn't the result of a hardware error) and yet the programmer is unable to explain how the algorithm produced its output" (2019, p. 9). (There's that tricky little word again.)

Both in [INTRO] as in [MYTHS] we stressed how AI recapitulates the history of philosophical thought (*nihil novum sub sole*…), and lends itself as a stage on which to recast the metaphysical dilemmas of ages past. We see that extremely clearly in this case, where the topic of machine creativity reanimates the old debate about free will and determinism, just like Samuel stated earlier. Of course, we lack the space to do justice to such a deep human pondering, but let us note that an astute scientific innovator like Francis Galton thought it wiser to simply sidestep the issue (in a manner much reminiscent of Turing's own in regards to the objection from consciousness to machine thought, as was abundantly elaborated in [TESTS]):

> [W]e need not linger to re-open the unending argument whether man possesses any creative power of will at all, or whether his will is not also predetermined by blind forces or by intelligent agencies behind the veil, and whether the belief that man can act independently is more than a mere illusion. This matters little in practice, because men, whether fatalists or not, work with equal vigour whenever they perceive they have the power to act effectively. (Galton, 1906, p. 53)

The ontological underpinnings of behaviorism are worth bringing up if for no other reason than that Turing so often gets accused of being one (for a discussion of whether Turing should be considered a behaviorist see Proudfoot, 2017. Short answer: no). The most committed and consistent believer of this view of the world is in all likelihood Burrhus Frederic Skinner. Given that behaviorism, like positivism, Cartesianism and every other "ism" fallen out of fashion is customarily vilified by those who have scarcely read a word of it, I feel it justified to quote Skinner's own views on the subject of personal creativity. In this case, in a lecture in which he compares 'having a poem' with 'having a baby'[63]:

> A person produces a poem and a woman produces a baby, and we call the person a poet and the woman a mother. Both are essential as loci in which vestiges of the past come together in certain combinations. The process is creative in the sense that the products are new. […] What is threatened, of course, is the autonomy of the poet. The autonomous is the uncaused, and the uncaused is miraculous, and the miraculous is God. For the second time in a little more than a century a theory of selection by consequences is threatening a traditional belief in a creative mind. And is it not rather strange that although we have abandoned that belief with respect to the creation of the world, we fight so desperately to preserve it with respect to the creation of a poem? (Skinner, 1971, p. 354)

What are we actually saying when we claim that someone has originated something? Making sense of this conundrum would demand of us the disentangling of several thorny subtleties, for we speak with words that we have not created, and yet we feel them intimately ours. Many have claimed that it is nothing but our ignorance which allows us to hang on to the gall of claiming anything we say as original. George Eliot said: "One couldn't carry on life comfortably without a little blindness to the fact that everything has been said better than we can put it ourselves" (1996, p. 146) and Oliver Wendell Holmes affirmed that "honest thinkers are always stealing unconsciously from each other. Our minds are full of waifs and strays which we think are our own. Innocent plagiarism turns up everywhere" (cited in Leary, 2006, p. 37). At an even deeper level, we are all reliant on extant language, which is part of that immeasurable boon that Samuel Ichiye Hayakawa called the "free gifts from the dead" (1941, p. 21).

If that is so in our case, then to an even greater extent when machines address us from the pages of the volumes on AI they speak with a mighty voice culled from the successive pruning maneuvers of accomplished stockbreeders of prose. Simply put, we get the *crème de la crème*, the cream of the crop with the cherry on top. It is not devoid of irony that in order to pursue the example more clearly we

too must play the part in the stigmergic dance, and propagate once more those yummiest bites of yore. We could hardly do better than offer one of the most beloved passages out of *The Policeman's Beard Is Half Constructed*, a book advertised as the first to be written by a computer program, allegedly by the program Racter (short for *Raconteur*, French for 'storyteller'):

> Love is the question and the subject of this essay. We will commence with a question: does steak love lettuce? This question is implacably hard and inevitably difficult to answer. Here is a question: does an electron love a proton, or does it love a neutron? Here is a question: does a man love a woman or, to be specific, does Bill love Diane? The interesting and critical response to this question is: no! He is obsessed and infatuated with her. He is loony and crazy about her. That is not love of steak and lettuce, of electron and proton and neutron. This dissertation will show that the love of a man and a woman is not the love of steak and lettuce. Love is interesting to me and fascinating to you, but it is painful to Bill and Diane. That is love! (Racter & Chamberlain, 1984, p. 9; Swirski, 2013, p. 31; Hofstadter, 1995, p. 473)

Over the ensuing decades many readers must have been left nonplussed wondering, I hardly doubt it, whether their own loves be as noble as that between lettuce and steak. In the face of such a display of semantic virtuosity Douglas Hofstadter raised a healthy dose of skepticism regarding what had been the machine's actual contribution and what that of its human editors, especially William Chamberlain, Racter's programmer:

> Obviously, the passages by Racter quoted by me were culled by Chamberlain and friends from huge reams of output from Racter over a period of years. Moreover, you are also seeing just a little bit from the book, thus a double selection process has taken place - part by them, part by me. You are thus being exposed to the very choicest bits of all! What if we were instead allowed to see unfiltered, uncensored Racter output? We would probably be less impressed. (Hofstadter, 1995, p. 475)

Hofstadter then goes on to add a further note of caution, which is so clear and to the point that it merits being related in full, regarding the artistic products of computers: vendors are wary to tell us how the (metaphorical) sausage is made, which is never a good sign:

> In my experience, when computer prose, art, or music is exhibited, especially in the popular press, very little information is generally provided about how it was made. Why is that? Without such information, of what meaning is the exhibit?
>
> Suppose someone showed you a brilliant essay on humor and told you it had been "written by a computer". If subsequently you found out that it had been plucked whole from Arthur Koestler's book

> *The Act of Creation*, you would surely feel defrauded. It would make little difference if you were further informed that Koestler's entire book, along with a hundred million other books on all sorts of topics, had been stored inside the computer, and that a program had selected this particular book and from it this particular passage, and printed it out. It would still be no different from plagiarism - clever plagiarism, perhaps, but that's all.
>
> Now consider a somewhat more complex scenario. The same brilliant passage by Koestler is chopped up into a series of ten-word segments, and for each word in each segment, its part of speech is supplied. All this information is loaded into the computer's memory. In addition, a sophisticated grammar of English is given to the computer, along with overall instructions to try to arrange all the ten-word segments into one long passage that is grammatical from start to finish. The program runs for a good while and comes up with a candidate passage. It is not identical with Koestler's piece, but the two documents do have several stretches that are identical for a few hundred words in a row. Then the program is slightly modified and run again. This time, *mirabile dictu*, it comes up with Koestler's piece in its entirety. Does this piece of computer-generated prose deserve exhibition? Perhaps, but it would certainly be misleading to claim that a computer had "written" the passage, because the word "write" connotes the making of something from scratch. (Hofstadter, 1995, p. 480)

This idea of a machine merely shuffling and recombining snippets of delicate human-generated text brings us to the notion of machine intelligence as consuming the 'fossil fuel' of extant, archived human thought. Up until now, as in the case of ELIZA (see [MYTHS] and [TESTS]), machines that seemingly think are "thinking" by borrowing the machinery of our own brains. It is we who are endowing them with the power of our own thoughts, by believing that something substantive lurks beneath their text output. For a related example, Noam Chomsky's "colorless green ideas sleep furiously" is a string that has mutated, it has acquired meaning even though it initially had none, by the slow process of accretion after so many students were exposed to it while studying linguistics. It has borrowed it from our collectively thinking it. These things that we run, run on the machinery of our brains.

Particularly in the case of the currently popular and successful arrival of deep learning neural nets, we can understand AI as using up the fossil fuel of predigested human thought. Just like when we burn oil we are consuming the energy that the Sun deposited there over millions of years, or like a prodigal bon-vivant may quickly consume the riches his grandfather laboriously amassed. This is what happens, for instance, in modern automatic translation engines, which rely on gigantic corpora of translations made by human experts. The analogy of Big Data to a scarce natural resource is gaining in popularity

day by day: "In the business world, data is increasingly framed as an economic asset of critical importance, a commodity on a par with scarce natural resources" (Puschmann & Burgess, 2014, p. 44).

> The trouble with this sort of purely statistical machine learning is that it depends on having enormous amounts of data, and data predigested by human brains. Computers can recognize Internet images only because millions of real people have reduced the unbelievably complex information at their retinas to a highly stylized, constrained, and simplified Instagram of their cute kitty, and have clearly labeled that image too. The dystopian fantasy is simple fact: We're all actually serving Google's computers, under the anesthetizing illusion that we're just having fun with LOLcats. And yet even with all that help, machines still need enormous data sets and extremely complex computations to be able to look at a new picture and say, "kitty-cat!"—something babies can do with just a few examples. (Gopnik, 2015, p. 139)

For a cute and simple approximate example of this process, we may consider the following poemlet from a now long defunct website of Darwinian poetry, that while claiming no provenance from exclusively mechanical machinations does share a kinship with those when it comes to its method of composition:

Darwinian Poetry *# 17717*

you lie beautiful
beating beyond love
imagine
beneath chairs
dark stars magic
everything frozen strangely in

you

The way the website worked was by chopping up thousands of famous celebrated poems into strings of about two or three words each, which were then recombined at random. Then the resulting entities were made to compete with each other by being placed side by side and voted for by the visitors to

the site. The losing poems were killed off, and the winners got to keep on living and even 'mate', a process in which two successful poems would have some of their strands exchanged with one another. Yet it is not altogether completely different a process from what currently semi-coherent text-generators like GTP-2 are accomplishing, in many of which, if we are to generously indulge a big heap of willing suspension of disbelief, oftentimes a semblance of sense-making is to be found.

Now, when it comes to creativity in game-playing, we must remark that dazzled by the otherworldly play of the world's top AI programs, it is easy to lose sight of an often critical ingredient in their development: the very human stories and drives of their creators and their adversaries. "Contrary to the naïve conception of the autonomy of artificial intelligence, in the architecture of neural networks many elements are still deeply affected by human intervention" says media theorist Matteo Pasquinelli (2017, p. 7). The gleaming end result may lead us to forget how vital this strand of human passion was in the creation of their code. We see fascinating examples of this in the stories of Checkers and Go.

In the case of Checkers, computer scientist Jonathan Schaeffer became completely obsessed with having his program, Chinook, defeat Marion Tinsley, an intriguing figure universally believed in the Checkers community to be the best player who ever lived, a man who was in part "almost like an artificial intelligence—narrow but extraordinarily capable" (Madrigal, 2017, ¶ 31). Interviewed by a newspaper before their match, Tinsley had declared: "I can win. I have a better programmer than Chinook. His was Jonathan, mine was the Lord" (Schaeffer, 1997, p. 285). As Alexis Madrigal relates it in his fascinating account of the fruitful rivalry, while the match was hailed as Man vs. Machine, "the quick wits of a human versus the brute computing power of a supercomputer" (¶ 6), in actuality both contenders agreed that "this was a battle between two men, each having prepared and tuned a unique instrument to defeat the other" (¶ 6).

When it comes to Go, it is easy to become entranced by the lightning speed at which AlphaGo (a program the logic of whose inner workings, in its rebirth as AlphaZero, we already described in [GAMES]) taught itself to play surpassing the level of the reigning champions of the world. What is less known is that, before going on to defeat 18-time world champion Lee Sedol, AlphaGo had sharpened its claws by playing European master Fan Hui and defeating him in five games out of five. Marcus du Sautoy relates that "Fan Hui credits his matches with AlphaGo with teaching him new insights into how to play the game. In the following months his ranking went from 633 to the 300s"

(2019, p. 19). Where the story turns fascinating, however, is in the role that Fan Hui would then play in the evolution of AlphaGo. After his defeat, Hui's loss had been mercilessly dismissed as a meaningless achievement for AlphaGo. Therefore, when the DeepMind team wanted someone who could test the program for any entrenched weaknesses, he was eager to hop on. "Perhaps a bit of him felt that if he could help make AlphaGo good enough to beat Sedol, it would make his defeat less humiliating" suspects du Sautoy (p. 19).

> As Fan Hui played he could see that AlphaGo was extremely strong in some areas but he managed to reveal a weakness that the team was not aware of. There were certain configurations in which it seemed to completely fail to assess who had control of the game, often becoming totally delusional that it was winning when the opposite was true. If Sedol tapped into this weakness, AlphaGo wouldn't just lose, it would appear extremely stupid. The DeepMind team worked around the clock trying to fix this blind spot. (du Sautoy, 2019, p. 20)

I certainly cannot blame those who overstate the claims of the machines, as I myself have only felt too vividly the temptation of leaving the human element out of the loop when telling of the feats of artificial systems. A personal anecdote may better drive the point home. As I write these lines a variation on Johann Pachelbel's *Canon in D Major*, one of my favorite pieces, plays in the background. For over two decades I have collected as many spins on Pachelbel's simple tune as I could get my hands on, a task that has been considerably eased by the fact that its appealing melody has delighted musicians all over the world and made its way into countless other songs over the centuries. But the one I'm listening to right now was written by no human, it was generated by OpenAI's MuseNet[64], out of the first seven seconds of Pachelbel's original Canon, which I fed it as a MIDI file. However, unless specifically prompted to give more detail, I prefer to omit the fact that the rest of the piece was composed in short bursts of six new seconds of music or so, at every which juncture I was given the option of choosing from between four different alternatives which one the program was to carry forward (which ended up being a painful choice in those cases in which two directions were at once widely different and very promising). Every new click made the piece grow and I felt as if Rumpelstiltskin were silently weaving his tapestry of gold (if only GPT-2 were half as good in generating sensible prose, I wished in vain, as I struggled to put the finishing touches on the thesis). And yet, for all my enjoying of the new piece (and despite my every supposition on the matter), I felt an underlying, bittersweet note, as if all the melody-cranking of this genie-in-a-box was only working at the cost of cheapening the human experience of music-making.

Troubled and amazed by the quality of the ensuing composition and how it would have certainly passed under my radar as human-made had I not known the truth, I sent it to a friend who is an extremely accomplished pianist without telling her what it was, and asked for her opinion about it. When I read her reply a wave of relief washed over me:

It seems to have been made by a computer. It lacks phrasings and intensities, and you can't really appreciate the counterpoint. It sounds so artificial that I don't know what else to tell you. Almost as if played by a street organ. Stick to the original, you have no need for variations on a theme so sublime and well wrought.

References

du Sautoy, M. (2019). *The Creativity Code: How AI is Learning to Write, Paint and Think*. London: 4th Estate.

Eliot, G. (1996). *Daniel Deronda*. Kent, UK: Wordsworth Editions.

Galton, F. (1906). Eugenics as a Factor in Religion. In *Sociological Papers*, Volume II. London: Macmillan & Co.

Gopnik, A. (2015). Can Machines Ever Be As Smart As Three-Year-Olds? In J. Brockman (Ed.), *What to Think About Machines That Think*. NY: HarperCollins.

Hayakawa, S.I. (1941). *Language in Action*. NY: Harcourt, Brance and Jovanovich.

Hofstadter, D. (1995). On Computers, Creativity, Credit, Brain Mechanisms, and the Turing Test. In D. Hofstadter and the Fluid Analogies Research Group (Eds.) *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. NY: Basic Books.

Hofstadter, D. (2000). *Gödel, Escher, Bach: An Eternal Golden Braid*. London: Penguin. (Original work published 1979).

Leary, D. (2006). The Missing Person in the Conversation: Oliver Wendell Holmes, Sr., and the Dialogical Self. *International Journal for Dialogical Science* 1, no. 1 (2006): 33-39.

Madrigal, A. (2017, July 19). How Checkers Was Solved. *The Atlantic*. Retrieved from: https://www.theatlantic.com/technology/archive/2017/07/marion-tinsley-checkers/534111/

Pasquinelli, M. (2017). Machines that Morph Logic: Neural Networks and the Distorted Automation of Intelligence as Statistical Inference. *Glass Bead*, 1: "Logic Gate: The Politics of the Artifactual Mind". Retrieved from: www.glass-bead.org/article/960

Proudfoot, D. (2017). Turing's Concept of Intelligence. In B.J. Copeland, J. Bowen, M. Sprevak, & R. Wilson (Eds.) *The Turing Guide*. Oxford: Oxford University Press, pp. 301–307.

Puschmann, C. & Burgess, J. (2014). The Politics of Twitter Data. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.) *Twitter and Society*. NY: Peter Lang, pp. 43–54.

Racter & Chamberlain, W. (1984). *The Policeman's Beard Is Half Constructed*. NY: Warner Books.

Schaeffer, J. (1997). *One Jump Ahead: Challenging Human Supremacy in Checkers*. NY: Springer.

Shannon, C. (1956). A Chess-Playing Machine. In J.R. Newman (Ed.) *The World of Mathematics*. NY: Simon & Schuster, p. 2124-2133.

Skinner, B.F. (1971). A Lecture on "Having" a Poem. In B.F. Skinner (Ed.) *Cumulative Record: A Selection of Papers*. (Third Edition). NY: Appleton-Century-Crofts, pp. 345–358.

Swirski, P. (2013). *From Literature to Biterature: Lem, Turing, Darwin, and Explorations in Computer Literature, Philosophy of Mind, and Cultural Evolution*. Canada: McGill-Queen's University Press.

Turing, A.M. (1948). Intelligent Machinery. In B.J. Copeland (Ed.) *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press, pp. 410–432.

Wittgenstein, L. (1998). *Culture and Value: A Selection from the Posthumous Remains*. [Edited by Georg Henrik von Wright in collaboration with Heikki Nyman]. Oxford, UK: Blackwell Publishers.

Endnotes

[1] And as for the embarrassment that may afflict he who fails to live up to the citations he has summoned, well, as so often is the case, the past has already expressed it more grandiosely and eloquently, like the brothers James and Horace Smith do here when parodying the style of Samuel Johnson (in words that were later playfully taken by Edgar Allan Poe to serve as the epigraph for one of his own pieces):

"A swelling opening is too often succeeded by an insignificant conclusion. Parturient mountains have now produced muscipular abortions; and the auditor who compares incipient grandeur with final vulgarity is reminded of the pious hawkers of Constantinople, who solemnly perambulate her streets, exclaiming, "In the name of the Prophet—figs!" " (Smith & Smith, 1879, p. 33)

[2] Charting the computer's precise lineage can lend itself to fractious disputes. But it can also be a source of fun, of insight or of both:

"It is sometimes said in jest that if Turing was the father of the computer, von Neumann was the obstetrician or midwife. Quite obviously a third element is missing: the womb. This, it needs at last to be recognized, was McCulloch's machine." (Dupuy, 2000, p. 66)

[3] In fact, that very perfectionism may have cost Babbage dearly, as his ambition of building an enormously complex difference engine, instead of a smaller but practical one that could be sold right away, probably delayed the widespread adoption of his ideas a whole century (Bowden, 1953, p. 15).

[4] For a highly poetic rendering of our all too human tendency to liken the mind to anything but itself, including mirrors, consider the following passage by George Eliot, that crown jewel of psychological *belles lettres*:

"It is astonishing what a different result one gets by changing the metaphor! Once call the brain an intellectual stomach, and one's ingenious conception of the classics and geometry as ploughs and harrows seems to settle nothing. But then, it is open to someone else to follow great authorities and call the mind a sheet of white paper or a mirror, in which case one's knowledge of the digestive process becomes quite irrelevant. It was doubtless an ingenious idea to call the camel the ship of the desert, but it would hardly lead one far in training that useful beast. O Aristotle! if you had the advantage of being "the freshest modern" instead of the greatest ancient, would you not have mingled your praise of metaphorical speech as a sign of high intelligence, with a lamentation that intelligence so rarely shows itself in speech without metaphor,—that we can so seldom declare what a thing is, except by saying it is something else?" (Eliot, 1997, p. 125)

For an insightful in-depth treatment of the theoretical consequences of modeling the mind as a computer see Hurtado, 2017.

[5] His friend T.S. Eliot once described him (in a private letter) as "a great wonderful fat toad bloated with wisdom." (Eliot, 2011, p. 108)

[6] Butler's closing remarks in the same piece (though it is hard to discern whether they be not at least partially tongue-in-cheek) radiate such passionate neo-luddite appeal that they might well have inspired Frank Herbert (1965), one of science-fiction's most dearly cherished authors, in his masterpiece of geopolitical and philosophical intrigue, *Dune*, to give the name 'Butlerian Jihad' to a crusade that led to a galaxy-wide ban on thinking machines:

"Day by day, however, the machines are gaining ground upon us; day by day we are becoming more subservient to them; more men are daily bound down as slaves to tend them, more men are daily devoting the energies of their whole lives to the development of mechanical life. The upshot is simply a question of time, but that the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question. Our opinion is that war to the death should be instantly proclaimed against them. Every machine of every sort should be destroyed by the well-wisher of his species. Let there be no exceptions made, no quarter shown; let us at once go back to the primeval condition of the race. If it be urged that this is impossible under the present condition of human affairs, this at once proves that the mischief is already done, that our servitude has commenced in good earnest, that we have raised a race of beings whom it is beyond our power to destroy, and that we are not only enslaved but are absolutely acquiescent in our bondage." (Butler, 1863, ¶ 7)

[7] Just as in Wiener's, in the following passage from William James we see how the single-mindedness of machines can coexist with their endowment with minds as a cause for concern:

"A machine in working order functions fatally in one way. Our consciousness calls this the right way. Take out a valve, throw a wheel out of gear or bend a pivot, and it becomes a different machine, functioning just as fatally in another way which we call the wrong way. But the machine itself knows nothing of wrong or right: matter has no ideals to pursue. A locomotive will carry its train through an open drawbridge as cheerfully as to any other destination." (1879, ¶ 37)

[8] Also in psychotherapy, as is well illustrated by the following example, dealing with *personal styles* among experienced practitioners and the difficulties facing disciples who seek to acquire the master's way: A famed and reputedly brilliant clinical psychologist had successfully dealt with a chronically depressed patient by—during her most heightened crises—attentively listening to her and then, matter-of-factly but looking her straight in the eye, saying: "Well, then go ahead and kill yourself!". These ritual words had always succeeded in putting the patient at ease and making her see things in a sobering perspective. The therapist was understandably aghast, then, when upon returning from a long vacation she came to learn that the student in training under whose care she had temporarily left the patient had been only too keen to echo her enchantment, and the patient, in turn, had this time obediently heeded the advice.

[9] In a symmetrical way, many qualitative researchers have, for similar reasons, adopted the techniques of their quantitative colleagues. See Musa, Olivares & Cornejo, 2015.

[10] It bears mentioning that in a volume put forth by Edge Magazine, attempting to capture the thoughts of nearly two hundred scholars and thinkers on the topic of machines that think, Freeman Dyson offers the shortest response. After declaring his general skepticism that such machines will ever come to exist, he simply adds: "If I am wrong, as I often am, any thoughts I might have about the question are irrelevant. If I am right, then the whole question is irrelevant." (Dyson, 2015, p. 47)

[11] Although there are some differences in flavour and shading between the terms 'extropianism' and 'transhumanism' (as well as within the use of the term 'transhumanism' itself on the part of different writers) for the purposes of this essay we will use them interchangeably.

[12] In addition to the socialist antecedent, Burkhead (1997, ¶ 8) offers another biblical forebear to this grand scheme:

"The vision of a transhuman condition goes all the way back to Isaiah.

Never again will there be in it [the new Jerusalem]

an infant who lives but a few days,

or an old man who does not live out his years;

he who dies at a hundred

will be thought a mere youth;

he who fails to reach a hundred

will be considered accursed."

[13] It must be clarified, however, that despite the existence of certain foundational texts and certain prominent figures and institutions that act as attractors, there is no real unified organization that would encompass all of those that would identify as transhumanists. Speaking of AI makers, AI researchers and, for that matter, even transhumanists, as though they were one single unified front in terms of belief and purpose is a misleading overgeneralization. A cursory perusal of the individual writings of key figures will show just how manifold the viewpoints they hold are.

[14] Ever the masterful salesman, Kurzweil opens the article on his law with: "You will get $40 trillion just by reading this essay and understanding what it says" (2001, ¶ 2). Lest my own readers should abandon this paper and instantly flock there in pursuit of so tasty a reward, I must add, *malgré moi*, the spoiler that by the end of the piece he explains that: "The English word 'you' can be singular or plural. I meant it in the sense of 'all of you'" (2001, ¶ 268).

[15] Transhumanism critic HP LaLancette (2005) takes this form of reasoning to its paroxysmic logical conclusion, pointing out that the very same argument can also be used to prove that the end goal of natural selection is the creation of the toilet brush. All that is needed is to replace the relevant landmarks. Thus, the Big Bang took place 13,7 billion years ago, after which another 10 had to elapse for life on Earth to arise. The appearance of the digestive tract, however, took only a further 2,75 and from then on the sphincter showed up merely another 575 million years hence. This projection leads us to the inescapable conclusion: eventually the whole universe will turn into one giant toilet brush.

[16] Yudkowsky, who co-founded and is a research fellow at the Institute, describes its institutional mission thus:

"The mission of the Machine Intelligence Research Institute is to do today that research which, 30 years from now, people will desperately wish had begun 30 years earlier." (Horgan, 2016, ¶ 68)

[17] While not identical to theirs, this classification owes much clarity to Cave & Dihal's recent typology of the "ways in which these narratives [of hope and fear] could shape [AI] technology and its impact." (2019, p. 74)

[18] Not to mention that AI researchers do not merely consume sci-fi but produce it as well. To single but two prominent examples, both John McCarthy and Marvin Minsky, starring figures at the Dartmouth Conference on Artificial Intelligence, which many consider the official birthplace of the field (Kline, 2011), have contributed their talents to the narrative arts. Minsky co-authored the technothriller *The Turing Option* (Harrison & Minsky, 1992) and McCarthy (2014) penned the delightful short story *The Robot and the Baby*, which shows just how hard it is to prevent people from anthropomorphizing automata.

[19] In his Foreword to the Millennial Edition of *2001: A Space Odyssey*, Arthur C. Clarke reproduces a touching letter sent to him by astronaut Joseph Allen, mission specialist on the Space shuttle program:

"Dear Arthur, When I was a boy, you infected me with both the writing bug *and* the space bug, but neglected to tell me how difficult either undertaking can be." (Clarke, 2000, p. xviii)

[20] Carnegie Mellon (academic home of Newell and Simon) is not just any university when it comes to the history of AI. Along with Minsky's MIT, McCarthy's Stanford and the Stanford Research Institute, it is one of the main four centers where AI took off. Seeking to characterize their differing styles, Pamela McCorduck offered this droll analogy between AI and the garment industry:

"Consider MIT haute couture, the Women's Wear Daily of the field. No sooner do hemlines go down with enormous fanfare than they go up again, the provinces growing dizzy with trying to keep pace and usually falling behind. MIT thinks itself stylish, but outsiders have been known to call it faddish. Carnegie Mellon, on the contrary, represents old-world craftsmanship, attending to detail and using the finest materials. These qualities presumably speak for themselves in gowns you can wear to a dinner party ten years from now and never fear the seams might part. But classic can be stodgy: if

Queen Elizabeth of England bought artificial intelligence, she'd surely buy at Carnegie Mellon. Stanford has two ateliers. The first is the Levis' jeans of AI: sturdy, durable, democratic; worn by socialites and welfare clients alike; and mentioned proudly by everyone in the trade whenever questions of practicality or utility come up. The other is Nudist World, incorporating After Six; this shop is visionary about the formal wear of the future, but meanwhile remains naked. Finally, Stanford Research Institute is Seventh Avenue. Maybe those models are knock-offs, but hardly anyone can afford haute couture, and except for the jeans people, who else is going to bring AI into the real world?" (McCorduck, 1979, p. 112)

[21] Renowned, among other things, for being the namesake and coiner of Sturgeon's Law, which states that while it's true that 90% of science fiction is crap, that is only because 90% of *everything* is crap.

[22] Compare with Dryden's (1913) rendering of Pygmalion's enthrallment to his creation, as told by Ovid:

> *Pleas'd with his Idol, he commends, admires,*
>
> *Adores; and last, the Thing ador'd, desires.*

[23] The three laws made their first formal appearance in Asimov's (1942) short story *Runaround*. To this story, Marvin Minsky claims a deep debt: "After 'Runaround' appeared in the March 1942 issue of Astounding, I never stopped thinking about how minds might work. Surely we'd someday build robots that think. But how would they think and about what?" (Minsky cited in Markoff, 1992, ¶18).

[24] Contrary to what the example suggests, the goal of some AI system needs not be particularly stupid to be extremely dangerous. Stephen Omohundro has argued that even a chess-playing robot "will indeed be dangerous unless it is designed very carefully. Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems." (2008, p. 483)

[25] If you can't beat them, join them, folksy wisdom asserts, and that is precisely what Yudkowsky did from 2010 to 2015 when he wrote his acclaimed spin on the Harry Potter franchise (Yudkowsky, 2015). Hailed as one of the most successful fan fictions ever written (Whelan, 2015), *Harry Potter and the Methods of Rationality* portrays Harry as a precocious genius that unleashes the whole arsenal of scientific reasoning upon the functioning of the magic world in order to maximize his own power (and optimize the world while he's at it). Keeping in line with Geraci's (2012, p. 40) claim that the incursions of the AI community into the realm of fiction crafting are more often than not evangelical in nature and are never written just for fun, *HPMOR*, as it is popularly known, is an attempt, much like Yudkowsky's *Center for Applied Rationality*, to induct young talents into the practice of Bayesian thinking, which could set them on a path of preventing the emergence of hostile superintelligences.

[26] We mean 'tacit' in the sense of Polanyi, 1983.

[27] A very vivid case in point is a recent flashy headline that made the rounds of social media, to the effect that Facebook had been forced to shut down some of its Artificial Intelligence agents since they had developed their own secret language and started communicating with each other to the befuddlement of their creators (Griffin, 2017; Bradley, 2017; Collins, 2017). What actually happened, though, is that chatbots designed for interaction with humans in a negotiation setting drifted from using conventional English and the researchers simply refined their reward schema to keep them on track with language that was grammatical (Lewis et al., 2017).

[28] In fact, "Can a Machine Think" became the unofficial name for *Computing Machinery and Intelligence*, after James R. Newman thus re-titled it in his anthology *The World of Mathematics* (Halpern, 2006). Interestingly, Newman also chose to preface Turing's paper with this bit from Emerson: "Beware when the great God lets loose a thinker on this planet" (Newman, 1956, p. 2088). Even more interesting, the only other epigraph from Emerson in the anthology precedes Claude Shannon's paper *A Chess-Playing Machine* and reads "Things are in the saddle and ride mankind" (Newman, 1956, p. 2124).

[29] The quotations correspond to McPherson, 1950, p. 545 and Turing, 1950, p. 433, respectively.

[30] But let us not be wholly disheartened; we should rather find comfort in Seneca's timeless wisdom:

"The earlier writers have not, I think, exhausted the possibilities; rather, they have opened up the way. It makes a big difference whether you take up a spent subject or one that has merely been treated before. A topic grows over time; invention does not preclude inventiveness. Besides, the last to come has the best of it: the words are all laid out for him, but a different arrangement lends them a fresh appearance." (2015, p. 258)

[31] While this would seem to place him as the father of Artificial Intelligence as a field, there is a sense in which he isn't, at least insofar as the American branch of AI is concerned. Here is Pamela McCorduck on the Dartmouth Conference:

"In a logical genealogy, Turing would be central. He held what were to be some of the central ideas of AI—the symbolic nature of the computer, the necessity to look at comparable functions instead of comparable hardware in humans and machines—very early. But the history of ideas has its own way of doing things and, as it happened, Turing's work had practically no influence on most people at the Dartmouth Conference. For instance, Minsky felt himself much more influenced by McCulloch and Shannon (especially Shannon's early chess paper); Simon considered Turing of no particular influence on his work." (McCorduck, 2004, p. 113)

[32] I employ the term 'criterion' to sidestep the controversy over whether Turing offered a definition proper of intelligence. Copeland (2000, p. 434) is adamant that Turing never offered an "operational definition" and that those who claim so, like Block (1990) French (2000) and even his biographer, Hodges (1992), are mistaken.

[33] Much has been made of the example Turing uses in order to introduce his imitation game, that of a man imitating a woman, especially by feminist scholars (Sack, 1996). I will not comment on its possible significance so as not to overburden an already convoluted discussion.

[34] Compare this to La Mettrie's analogous rejoinder, delivered two hundred years before to the religious objection to his doctrine that pure matter could think and feel without need of an immaterial soul: "it is irreligious to limit the Creator's supreme power by claiming that he cannot have made matter think - he who with a single word created light" (1750/1996, p. 65). In one of those capricious coincidences that link great men by happenstance, it should be noted that both natural philosophers, that so advanced the idea of thinking machines, died a mere handful of days before turning forty-two because of something they ate. Here the differences end, however, for La Mettrie, epicure that he was, incurred a gastric illness by gorging on pheasant pâté, while Turing bit into a cyanide-laced apple, which many commentators have interpreted as his taking his own life after the abuse he suffered at the hands of the British legal system.

[35] In 1958 he would sign, along Bertrand Russell and several other luminaries, a public letter arguing for homosexuality to be decriminalized in British law (Annan et al., 1958).

[36] While not being completely equivalent to it, our proposal is in some ways analogous to that of Brian Christian's admirable book *The Most Human Human* (2011):

"Here's the thing: beyond its use as a technological benchmark, beyond even the philosophical, biological, and moral questions it poses, the Turing test is, at bottom, about the act of communication. I see its deepest questions as practical ones: How do we connect meaningfully with each other, as meaningfully as possible, within the limits of language and time? How does empathy work? What is the process by which someone comes into our life and comes to mean something to us? These, to me, are the test's most central questions—the most central questions of being human." (Christian, 2011, p. 14)

[37] Searle's (1980) target article in *Behavioral and Brain Sciences* featured no less than twenty-eight responses, many of which did not shirk from conjuring imaginings as fanciful (if not more so) than Searle's own, including among others Dr. Jekyll and Mr. Hyde, clockwork dancing dolls, neuron-inhabiting demons and a homunculi army in a robotic head. One even stated—in what has since then become a common running joke—that Searle's failure to comprehend Turing's point should not be considered his own fault, but be rather attributed to the Chinese homunculus living inside Searle's skull who is in charge of producing suitable English output out of the English input squiggles it receives (Abelson, 1980).

[38] In fact, Turing's proposal would pass with flying colors the bar exam of pragmatic philosophy as articulated by William James on his lectures on Pragmatism:

"It is astonishing to see how many philosophical disputes collapse into insignificance the moment you subject them to this simple test of tracing a concrete consequence. There can be no difference anywhere that doesn't make a difference elsewhere – no difference in abstract truth that doesn't express itself in a difference in concrete fact and in conduct consequent upon that fact, imposed on somebody, somehow, somewhere, and somewhen. The whole function of philosophy ought to be to find out what definite difference it will make to you and me, at definite instants of our life, if this world-formula or that world-formula be the true one." (James, 1922, p. 49)

[39] What novelist Richard Powers said to Bruno Latour during their conversation on the Turing Test seems particularly apropos here:

"[T]he problem of artificial intelligence is much like the problem of fiction, if you want to call fiction a problem. And the crisis of fiction is itself a Turing crisis; a crisis of correspondence, of free agents and the representations of agents, communicating through narrow bandwidths across great distances, urging us to take the simulation for the thing that it stands for." (Latour & Powers, 1998, p. 15)

[40] McCorduck herself defends the term as a "wonderfully appropriate name, connoting a link between art and science that as a field AI indeed represents" (2004, p. 115).

Massimo Negrotti (2002) has proposed the term 'naturoid' to introduce an important conceptual distinction: while all conventional technology can be considered artificial in the lax etymological sense of being "made through art" the use of 'artificial' in artificial intelligence refers specifically to the attempt to imitate via technology something already existing in nature.

[41] This is reminiscent of William James' definition of *thing*, which is even wider in its non-boundedness:

"But what are things? Nothing, as we shall abundantly see, but special groups of sensible qualities, which happen practically or aesthetically to interest us, to which we therefore give substantive names, and which we exalt to this exclusive status of independence and dignity. But in itself, apart from my interest, a particular dust-wreath on a windy day is just as much of an individual thing, and just as much or as little deserves an individual name, as my own body does." (1890, p. 285)

[42] The most famous (or infamous, depending on who you ask) attempt to implement an actual Turing Test is the Loebner Prize competition, which has drawn heavy fire from the AI community. Dennett, formerly chairman of the competition, in the end resigned because of a difference of visions regarding its goals and implementation (Dennett, 2004, p. 315), Marvin Minsky once offered a prize of $100 to the person who could get Hugh Loebner to stop holding it so that the AI community should be spared "the horror of this obnoxious and unproductive annual publicity campaign" (Barrat, 2013, p. 49) and Hofstadter said that "what is needed is a prize for advances in basic research, not a prize for window-dressing" (1995, p. 491).

The late Loebner, for his own part, seems to have been largely unconcerned with the scholastic hair-splitting of formal AI theory but wanted simply to do his part to spur technological advancement toward an age of abundance and be remembered for it, to boot (Sundman, 2003). In a reply to Stuart Shieber (1994), who criticized the way the contest was being run and suggested alternative uses for the prize money, he wrote (after declaring that he was preventing no one else from doing their own thing and that he eagerly looked forward to The Shieber Prize) that:

"It amuses me to imagine a day in the distant future when humans have become extinct, surpassed by our creations, robots, who roam the universe. I like to think that these robots may have a memory of us humans, perhaps as semi-mythic fractious demigods from the distant past who created them. And, just possibly, they will remember me." (Loebner, 1994, ¶30)

[43] Some overhyped claims, such as the announcement that Eugene Goostman, a chatbot that was passed off as a 13-year-old Ukrainian boy, had been the first program to beat the Turing Test, undermine the deep aspect of Turing's proposal, reducing it to little more than a debate about how easily human observers can be deceived by clever tricks.

[44] Or was it "red on yellow, friendly fellow; red on black, deadly attack"? The point is that if you were truly facing the Serpent Screening, the Asp Assessment or the Python Probe you should go about it really, really conscientiously.

[45] As per the cookbook of methodological rigor, the *cum grano salis* with which Rosenhan's study is to be taken is a tough pink nugget of coarse Himalayan rock salt. I am not accepting Rosenhan's interpretation for his alleged findings at face value, nor do I suggest my readers to do so, and even less so the many accounts thereof to be found in introductory psychology textbooks. However, even if marred by issues of design, his intention is fascinating and his paper well worth the read, especially if followed by the incisive critiques by Spitzer, 1975, 1976; Millon, 1975; Davis, 1976, and finally Rosenhan's (1976) thought-provoking rejoinder to some of these objections. Be that as it may, the scenario that Rosenhan brought forth to the literature is at the very least as worthy a *Gedankenexperiment* as Searle's and can be considered functionally equivalent to being moved to reflection by Ken Kesey's (1962) *One Flew Over the Cuckoo's Nest*, a novel very much in line with the same theme.

[46] For an exposition of Lipps' notion that we feel ourselves into the empathetically perceived, see Cornejo, 2016.

[47] The closing verses of *False Greatness* (Watts, 1762, p. 107), from his *Horae Lyricae*, reproduced here as originally published in 1762:

Were I ſo tall to reach the Pole,
Or graſp the Ocean with my Span,
I muſt be meaſur'd by my Soul :
The Mind's the Standard of the Man.

[48] I experienced a very acute illustration of this humbling reality when, on returning from a trip abroad, my brother informed me that he had finally made true his childhood dream of driving a luxury race car at breakneck speed through some winding roads in the Italian mountainside. As a part of the experience, the rental company offered to produce a DVD with video footage filmed from the driver's point of view, and this he invited me to see. Watching the rapidly scenery zoom by on that screen was positively terrifying. No matter how many times I looked to my left and told myself that my brother was sitting right by me and that was all the proof I needed that nothing had gone wrong during the drive, I could not dislodge from the depth of my stomach the dread that at any minute now the car would careen off the road to plunge down the hillside. Such is the power of our species' primeval acquaintance with the historical veracity of visual perception.

[49] It may bear saying that I'm not implying that we should treat every single interaction as a sort of Turing Test. Unless we had some reason to doubt someone's mindedness there seems to be no point in suspecting it a priori. That is also implied in Turing's reply to the Argument from Consciousness. Cornejo (2013, p. 248) has written a superb account of this disposition toward pre-reflexive trust:

"As often as we ordinarily swing between analytical and charitable dispositions, we swing likewise between trust and distrust toward others. The flow from one to the other is effortless and continuous, but our phenomenological experience varies fundamentally, depending on our disposition. Sometimes, the other is a minded subject, whose reasonability I take for granted in the same way that I am confident of the force of gravity on Earth. I am certain that she is right, and I do not require evidence to show the rightness of her statements. I simply follow her speech as I can follow melodies. Other times, the same other turns out to be an object of analysis, whose reasonability is questioned. In this case, I no longer am being with the other; I am following neither her dance nor her speech. I critically contemplate those actions before me and, by suspending my natural tendency to believe that they are right and reasonable, search for evidence that will allow me consider them as such. I am adopting a disposition to suspect, contrary to my natural disposition to trust. Human beings tend toward pre-reflexive trust. The most critical analyst of others' actions, who would question the rightness of every utterance, has to trust, at a given point, in something that the other says because doubt has an end. At the rock bottom of the understanding of others is a substratum of trust. Alternatively stated, the search for fundaments of our beliefs is finite. By contrast, epistemic doubt is endless because one will never gain logical certainty of the exact meaning of others' expressions. How, then, is human understanding possible if I can question the meaning of everyone's expressions? It is possible because, sooner or later, I trust in the rightness of what you are saying, hence exhibiting Davidson's 'charity principle'. Considered psychologically, because epistemic doubt originates from a disposition to suspect, the end of mistrust is rooted in the natural human disposition to empathize with others, in effect, to be-with-others."

[50] It has even been suggested that we owe the birth of the analytic detective novel to Edgar Allan Poe's attempting to figure out the mysterious mechanism of this piece of 'antebellum AI' (Grimstad, 2013).

[51] First formulated by Larry Tesler in the 70s, the original wording goes:

"Intelligence is whatever machines haven't done yet." (Tesler, 2019, ¶ 2)

[52] His could be a case of shared paternity, if we are to heed mathematician Stephen Wolfram's suspicions that Turing may have been the one to put in motion the idea:

"Turing had been working on a kind of statistical approach to cryptanalysis, and I am extremely curious to know whether Turing told Shannon about this and potentially launched the idea of information theory, which itself was first formulated for secret cryptographic purposes." (Wolfram, 2017, p. 44)

Also a putative father, a distraught Norbert Wiener bitterly blamed young and reckless Walter Pitts (of McCulloch-Pitts artificial neuron model fame) (McCulloch & Pitts, 1943) for losing a manuscript of his that would have given him priority over one of his competitors: Shannon (Kline, 2015, p. 10). Wiener seems to have been so worried about Shannon scooping him that he eventually avoided discussing his ideas with him altogether. As he confided to a family friend, he believed that Shannon was: "coming to pluck my brains" (Conway & Siegelman, 2005, p. 94).

[53] By Frederic Friedel, co-founder of ChessBase, and Mathias Feist. They ran into problems, however, getting the program generated from Turing's algorithm to make the exact same moves that Turing had written down in his record of one of the laborious and painstaking matches thus produced. Finally, Friedel consulted Donald Michie, who told him "You are trying to debug your program. You should debug Turing!" (in Copeland, 2017, p. 343). It was Turing who had made mistakes in following the steps of his own algorithm.

[54] See, in the case of virtual reality, Zyda, 2016

[55] The same principle was at play in the checkers-playing program of Arthur Samuel, coiner of the term 'machine learning':

"To speed up learning, Samuel would set up two copies of the programme, Alpha and Beta, on the same computer and leave them to play game after game with each other. The learning procedure consisted in the computer making small numerical changes to Alpha's ranking procedure, leaving Beta's unchanged, and then comparing Alpha's and Beta's performance over a few games. If Alpha played worse than Beta, these changes to the ranking procedure were discarded, but if Alpha played better than Beta then Beta's ranking procedure was replaced with Alpha's. As in biological evolution, the fitter survived. Over many such cycles of mutation and selection, the programme's quality of play increased markedly." (Copeland, 2004, p. 514)

[56] Here is a more detailed explanation of AlphaZero's learning process:

"[AlphaZero] began training from a clean slate with no chess knowledge other than the rules of the game […]. AlphaZero also did not use any available chess openings knowledge, and instead worked out its own openings as it trained and played against itself. […] During those nine hours, AlphaZero

played a total of 44 million games against itself – more than 1,000 games per second. At the same time, it continuously adjusted the parameters of its neural network so as to capture moves and outcomes from the most recent batch of games played against itself." (Sadler & Regan, 2019, p. 60)

[57] This observation was originally illustrated by science writer and cartoonist Randall Munroe (2013) in connection to a sparrow on his celebrated comic *XKCD*, but molecular data would seem to suggest that chickens are even more related of a species (Organ et al., 2008).

[58] Incidentally, when B.F. Skinner finally explained why he had never addressed Chomsky's attack on his theory, he declared: "In the first place, I should have had to read the review" (1971, p. 346).

[59] "The term reflects diverse meanings, contradictory uses, division on its academic worth, underdeveloped theoretical foundations, and a lack of standardized application. Its interpretation is often determined by the sector in which it is being used." (Todd, 2016, p. 4)

[60] Compare William James on the unshakeable might of habit:

"Habit is thus the enormous fly-wheel of society, its most precious conservative agent. It alone is what keeps us all within the bounds of ordinance, and saves the children of fortune from the envious uprisings of the poor. It alone prevents the hardest and most repulsive walks of life from being deserted by those brought up to tread therein. It keeps the fisherman and the deck-hand at sea through the winter; it holds the miner in his darkness, and nails the countryman to his log-cabin and his lonely farm through all the months of snow; it protects us from invasion by the natives of the desert and the frozen zone. It dooms us all to fight out the battle of life upon the lines of our nurture or our early choice, and to make the best of a pursuit that disagrees, because there is no other for which we are fitted, and it is too late to begin again." (James, 1890/1950, p. 121)

[61] This should go without saying, but quoting Kaczynski and being willing to engage critically with his ideas should by no stretch of the imagination be construed as offering even the tiniest shred of support for the criminal tactics he employed in pursuit of his goals. However, just as insists David Skrbina (2010) in his introduction to Kaczynski's papers, his ideas merit being read and pondered, not out of consideration for the writer, but out of consideration for our very future. I have written at length (see [TESTS]) about the ethical imperative of not allowing what I call 'overriding labels' to become superimposed atop an object of thought and forestall our own thoughtful engagement with it. For all that they might be tarnished with the "written by a terrorist and murderer" stamp, we only do a disservice to ourselves in refusing to consider them.

[62] This passage (with its idiosyncratic diacritical signs) is taken from Volume IV, Part IV, Chapter 6 of Eduardo Nolla and James T. Schleifer excellent (2010) historical-critical edition.

[63] A lecture which he wraps up, truth be told, in superb Skinnerian fashion, with the following words:

"And now my labor is over. I have had my lecture. I have no sense of fatherhood. If my genetic and personal histories had been different, I should have come into possession of a different lecture. If I deserve any credit at all, it is simply for having served as a place in which certain processes could take place. I shall interpret your polite applause in that light." (Skinner, 1971, p. 355)

[64] The neural network can be played with for free at https://openai.com/blog/musenet/