AI & SOCIETY (2020) 35:1009–1024
https://doi.org/10.1007/s00146-020-00966-4

OPEN FORUM

Echoes of myth and magic in the language of Artificial Intelligence

Roberto Musa Giuliano[1]

Received: 7 June 2019 / Accepted: 27 February 2020 / Published online: 7 April 2020

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

> *"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).*

1 Introduction

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"[1] (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:

✉ Roberto Musa Giuliano
rfmusa@uc.cl; robmusa@gmail.com

[1] School of Psychology, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile

[1] 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

> 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.

2 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 a related essay (Musa 2019) 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 emptied 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

Footnote 1 (continued)

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).

merely by contemporary literary theorists but also by distinguished pioneers from the field. Paramount among these is the founder of Cybernetics, Norbert Wiener[2] 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 an even more direct link 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

[2] The poet T.S. Eliot, a friend of his youth, once described him (in a private letter) as "a great wonderful fat toad bloated with wisdom." (Eliot 2011, p. 108).

> 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).[3]

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 2007; Cornejo and 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.

[3] 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)

> 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)[4]

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).[5] 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:[6]

> 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 impre-

[4] 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." (James 1879, ¶ 37)

[5] 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.

[6] In a symmetrical way, many qualitative researchers have, for similar reasons, adopted the techniques of their quantitative colleagues. See Musa et al. (2015).

> cise 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" (1990, 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. 422)

3 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[7]) 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 2012, 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

[7] 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)

> 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[8] 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…[9]

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.[10] 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

[8] 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.

[9] 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."*

[10] 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.

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).[11] 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).[12]

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) 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 poem by Brautigan (1967):

> 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

[11] 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).

[12] Transhumanism critic HP LaLancette (2007) 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.

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 full time to the preemptive forestalling of existential risk. 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 and 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, emphasis in the original)

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.

4 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 and Horvitz (2017, p. 4) observe that while AI and science fiction have always been associated, that association pointedly increased at the start of the 90 s. 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.[13] 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 et al. 2019; Fast and 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

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

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 has not been reared on the fantasies of Isaac Asimov and Arthur C. Clarke.[14] 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 HAL was a one-step alphabetical transliteration of IBM (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.[15] (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 and Mattei 2014; Tambe et al. 2008; Bates 2011). Robert Geraci noted the strength of the link when he visited the Robotics Institute of Carnegie Mellon University,[16] 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 2012, 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

[14] 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 and 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.

[15] 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)

[16] 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)

> 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 and 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 and 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 et al. 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[17], 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)[18]

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?[19] 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

[17] 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.

[18] 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.*

[19] 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).

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.[20] 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[21], 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" (Sotala 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" (Orwell 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,

[20] 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." (Omohundro 2008, p. 483)

[21] 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.

> 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 et al. (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 et al. 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" (Valsiner 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.

5 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 et al. 2011) of Artificial Intelligence, that is, a tacit commitment driving researchers in their quest to expand the discipline.[22] 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.[23] 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)

Funding This work is sponsored by grants from CONICYT and Pontificia Universidad Católica de Chile.

References

Asimov I (1942) Runaround. Astound Sci Fict 29:94–103

Barrat J (2013) Our final invention: artificial intelligence and the end of the human era. Thomas Dunne Books, Chicago

Barutta J, Cornejo C, Ibáñez A (2011) Theories and theorizers: a contextual approach to theories of cognition. Integr Psychol Behav Sci 45(2):223–246

Bates RA (2011) AI & SciFi: teaching writing, history, technology, literature, and ethics. In: Paper presented at 2011 ASEE annual conference & exposition, Vancouver, BC. https://peer.asee.org/17433. Accessed 25 Apr 2019

Bostrom N (2003) Ethical issues in advanced artificial intelligence. In: Smit I (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 University Press, Oxford, United Kingdom

[22] We mean 'tacit' in the sense of Polanyi 1983.

[23] 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).

Bostrom N (2015) It's still early days. In: Brockman J (ed) What to think about machines that think. HarperCollins, New Year, pp 126–127
Bova B (1974) The role of science fiction. In: Bretnor R (ed) Science fiction today and tomorrow. Harper & Row, New Year
Bradley T (2017) Facebook AI creates its own language in creepy preview of our potential future. Forbes. https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future. . Accessed 25 Apr 2019
Brautigan R (1967) All watched over by machines of loving grace. The Communication Company, San Francisco
Burkhead L (1997) Extropianism in the memetic ecosystem. Extropians Message Board
Butler S (1863) Darwin among the machines. Christchurch Press, June 13. http://www.nzetc.org/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html. Accessed 25 Apr 2019
Cave S, Dihal K (2019) Hopes and fears for intelligent machines in fiction and reality. Nat Mach Intell 1(2):74
Cave S, Coughlan K, Dihal K (2019) 'Scary robots': examining public responses to AI. In: Proc. AIES. http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_200.pdf. Accessed 25 Apr 2019
Chalmers DJ (2010) The singularity: a philosophical analysis. J Conscious Stud 17:7–65. http://consc.net/papers/singularity.pdf. Accessed 25 Apr 2019
Clark SRL (1995) Tools, machines, and marvels. In: Fellows R (ed) Philosophy and technology. Cambridge University Press, Cambridge
Clarke AC (2000) 2001: a space Odyssey. ROC, New Year
Collins T (2017) Facebook shuts down controversial chatbot experiment after AIs develop their own language to talk to each other. Daily Mail. https://www.dailymail.co.uk/sciencetech/article-4747914/Facebook-shuts-chatbots-make-language.html. Accessed 25 Apr 2019
Computer History Museum (2017) Oral History of John McCarthy [Video file]. http://www.youtube.com/watch?v=KuU82i3hi8c. Accessed 25 Apr 2019
Comrada N (1995) Golem and robot: a search for connections. J Fantas Arts 7(2/3):244–254
Cornejo C, Musa R (2017) The physiognomic unity of sign, word, and gesture. Behav Brain Sci 40:E51. https://doi.org/10.1017/S0140525X15002861
Cramer JG (1990) Technology fiction (Part I). Foresight update 8, March 15. http://www.islandone.org/Foresight/Updates/Update08/Update08.2.html. Accessed 25 Apr 2019
DeBaets AM (2015) Rapture of the Geeks: singularitarianism, feminism, and the yearning for transcendence. In: Mercer C, Trothen TJ (eds) Religion and transhumanism. Praeger, CA, pp 181–197
Deutsch D (2011) The beginning of infinity. Allen Lane, London
Dreyfus HL (1992) What computers still can't do: a critique of artificial reason. MIT, Cambridge, Mass
Dryden J (1913) The Poems of John Dryden, ed. by John Sargeaunt. Oxford University Press, London. https://www.bartleby.com/204/199.html. Accessed 25 Apr 2019
Dyson G (2005) Turing's cathedral. Edge. http://www.edge.org/conversation/george_dyson-turings-cathedral. Accessed 25 Apr 2019
Dyson F (2015) I could be wrong. In: Brockman J (ed) What to think about machines that think. HarperCollins, NY, pp 126–127
Egan G (1997) Diaspora. Orion, London
Eliot G (1997) The mill on the floss. Wordsworth Editions, Hertfortshire
Eliot TS (2011) [Letter written December 31, 1914 to Conrad Aiken]. In: The letters of T.S. Eliot, vol 1. London: Faber and Faber
Fast E, Horvitz E (2017) Long-term trends in the public perception of artificial intelligence. In: Thirty-first AAAI conference on artificial intelligence. https://arxiv.org/abs/1609.04904 (2016)
Fellows R (1995) Welcome to Wales: Searle on the computational theory of mind. In: Fellows R (ed) Philosophy and technology. Cambridge University Press, Cambridge
Feynman R (1985) Surely you're joking, Mr. Feynman! Adventures of a curious character. Bantam Books, New York
Foerst A (2004) God in the machine: what robots teach us about humanity and God. Dutton, New York
Geraci RM (2008) Apocalyptic AI: religion and the promise of artificial intelligence. J Am Acad Relig 76(1):138–166
Geraci RM (2012) Apocalyptic AI: visions of heaven in robotics, artificial intelligence, and virtual reality. Oxford University Press, Oxford
Goldsmith J, Mattei N (2014) Fiction as an introduction to computer science research. ACM TOCE 14(1):4
Good IJ (1966) Speculations concerning the first ultraintelligent machine. Adv Comput 6:31–88
Griffin A (2017). Facebook's artificial intelligence robots shut down after they start talking to each other in their own language. The Independent. https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html. Accessed 25 Apr 2019
Halpern M (2008) The Trojan Laptop. Vocabula review, vol 10, Issue 1. http://www.rules-of-the-game.com/com007-trojanlaptop.htm. Accessed 25 Apr 2019
Harrison H, Minsky M (1992) The Turing option. Warner, New Year
Herbert F (1965) Dune. Chilton books, Philadelphia
Herbert F (1974) Science fiction and a world in crisis. In: Bretnor R (ed) Science fiction today and tomorrow. Harper & Row, New Year
Hess DJ (1995) On low-tech cyborgs. In: Gray CH, Figueroa-Sarriera H, Mentor S (eds) The cyborg handbook. Routledge, New York, pp 371–378
Hurtado E (2017) Consequences of theoretically modeling the mind as a computer. Doctoral dissertation, Pontificia Universidad Católica de Chile. https://repositorio.uc.cl/handle/11534/21956
Ingold T (2007) Lines: a brief history. Routledge, Oxon
James W (1879) Are we automata? Mind 4:1–22. http://psychclassics.yorku.ca/James/automata.htm. Accessed 25 Apr 2019
Johnson G (1998) Science and religion: bridging the great divide. New York Times. http://www.nytimes.com/1998/06/30/science/essay-science-and-religion-bridging-the-great-divide.html. Accessed 25 Apr 2019
Kline R (2011) Cybernetics, automata studies, and the Dartmouth conference on artificial intelligence. IEEE Ann Hist Comput 33(4):5–16
Kress G (2010) Multimodality. Routledge, London
Kurzweil R (1990) The age of intelligent machines. MIT Press, Cambridge
Kurzweil R (2001) The law of accelerating returns. KurzweilAI.net. https://www.kurzweilai.net/the-law-of-accelerating-returns. Accessed 25 Apr 2019
Kurzweil R (2005) The singularity is near: when humans transcend biology. Penguin, New York
LaGrandeur K (2003) Magical code and coded magic: the persistence of occult ideas in modern gaming and computing. In: Paper presented at the conference of the society for literature, science and the arts. http://ieet.org/index.php/IEET/more/lagrandeur20131026. Accessed 25 Apr 2019
LaLancette HP (2007) The law of accelerating toilet brushes. http://blog.infeasible.org/2007/05/29/the-law-of-accelerating-toilet-brushes.aspx. Accessed 27 Nov 2012

Lancaster BL (1997) The golem as a transpersonal image: a marker of cultural change. Transpers Psychol Rev 1(3):5–11
Lancaster BL (2007) La esencia de la Kábala: La enseñanza interior del Judaísmo. EDAF, Madrid
Latour B (1987) Science in action. Harvard University Press, Cambridge, MA
Lewis M, Yarats D, Dauphin YN, 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. California Polytechnic State University, San Luis Obispo
Markoff, J. (1992, April 12). Technology; A Celebration of Isaac Asimov. *The New York Times*
McCarthy J (2014) The robot and the baby. In: Wilson DH, Adams JJ (eds) Robot uprisings. Simon & Schuster, London, pp 343–362. http://www-formal.stanford.edu/jmc/robotandbaby/robotandbaby.html. Accessed 25 Apr 2019
McCorduck P (1979) Machines who think. W. H. Freeman and Company, San Francisco
McCorduck P (2004) Foreword. In: Machines who think. A K Peters, Natick
McDermott D (1981) Artificial intelligence meets natural stupidity. In: Haugeland J (ed) Mind design. MIT, Cambridge, pp 143–160
Melzer A (2007) On the pedagogical motive for esoteric writing. J Polit 69(4):1015–1031
Minsky M (2007) Scientist on the set: an interview with Marvin Minsky/ interviewer: David G. Stork [Transcript]. https://web.archive.org/web/20071113031417/, http://mitpress.mit.edu/e-books/Hal/chap2/two3.html. Accessed 25 Apr 2019
Musa R (2019) Computer machinery and the benefit of the doubt. In: Myths, tests and games: cultural roots and current routes of artificial intelligence. Doctoral dissertation, Pontificia Universidad Católica de Chile. https://repositorio.uc.cl/handle/11534/28511
Musa R, Olivares H, Cornejo C (2015) Aesthetic aspects of the use of qualitative methods in psychological research. In: Marsico G, Ruggieri RA, Salvatore S (eds) Reflexivity and psychology. Information Age, Charlotte, pp 87–116
Newell A (1992) Fairy Tales. AI Mag 13(4):46–48
Noble DF (1999) The religion of technology: the divinity of man and the spirit of invention. Penguin, New York
Oatley K, Mar RA, Djikic M (in press) The psychology of fiction: present and future. In: Jaén EI, Simon J (eds) The cognition of literature. Yale University Press, New Haven
Omohundro SM (2008) The basic AI drives. In: Wang P, Goertzel B, Franklin S (eds) Artificial general intelligence 2008: proceedings of the first AGI conference. Frontiers in artificial intelligence and applications 171. Amsterdam: IOS, pp 483–492
Orwell G (1946) Politics and the English language. Horizon 13(76):252–265
Pels P (2013) Amazing stories: how science fiction sacralizes the secular. In: Stolow J (ed) Deus in machina: religion, technology, and the things in between. Fordham University Press, New York
Polanyi M (1983) The tacit dimension. Peter Smith Publisher Inc, Gloucester
Rosas R (1992) ¿Comerán los androides el fruto prohibido? Reflexiones acerca del Test de Turing. Apuntes de Ingeniería 45(1992):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 TN (1974) Science fiction as the imaginary experiment. In: Bretnor R (ed) Science fiction today and tomorrow. Harper & Row, New York
Shelley M (1818) Frankenstein; or, the modern prometheus. M. K. Joseph, London
Sotala K (2007). The logical fallacy of generalization from fictional evidence [Blog comment]. https://www.lesswrong.com/posts/rHBdcHGLJ7KvLJQPk/the-logical-fallacy-of-generalization-from-fictional#gchLRgHocaajGkEy2. Accessed 25 Apr 2019
Sturgeon T (1974) Science fiction, morals, and religion. In: Bretnor R (ed) Science fiction today and tomorrow. Harper & Row, New York
Tambe M, Balsamo A, Bowring E (2008) Using science fiction in teaching artificial intelligence. In: AAAI Spring symposium, pp 86–91
Taube M (1961) Computers and common sense: the myth of thinking machines. Columbia University Press, NY
Thiel P, Masters B (2014) Zero to one: notes on startups, or how to build the future. Crown Business, New York
Valsiner J (2009) Between fiction and reality: transforming the semiotic object. Sign Syst Stud 37(1/2):99–113
Van Leeuwen T (2004) Introducing social semiotics: an introductory textbook. Routledge, London
Vico G (1948) The new science of Giambattista Vico. (Translated by Thomas Goddard Bergin & Max Harold Fisch). Cornell University Press, Ithaca
Voegelin E (1952) The new science of politics. University of Chicago Press, Chicago
Whelan D (2015. The Harry Potter fan fiction author who wants to make everyone a little more rational. VICE. https://www.vice.com/en_us/article/gq84xy/theres-something-weird-happening-in-the-world-of-harry-potter-168. Accessed 25 Apr 2019
Wiener N (1964) God & Golem Inc: a comment on certain points where cybernetics impinges on religion. MIT, Cambridge
Williams R (1994) The metamorphosis of prime intellect. http://localroger.com/prime-intellect/mopiidx.html. Accessed 25 Apr 2019
Yudkowsky E (2004) Coherent extrapolated volition. Machine Intelligence Research Institute, Berkeley. https://intelligence.org/files/CEV.pdf. Accessed 25 Apr 2019
Yudkowsky E (2007a) The logical fallacy of generalization from fictional evidence. [Blog post]. http://www.lesswrong.com/posts/rHBdcHGLJ7KvLJQPk/the-logical-fallacy-of-generalization-from-fictional. Accessed 25 Apr 2019
Yudkowsky E (2007b) An alien god. [Blog post]. http://www.lesswrong.com/posts/pLRogvJLPPg6Mrvg4/an-alien-god. Accessed 25 Apr 2019
Yudkowsky E (2008a) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Ćirković MM (eds) Global catastrophic risks. Oxford University Press, New York, pp 308–345
Yudkowsky E (2008b) Cognitive biases potentially affecting judgment of global risks. In: Bostrom N, Ćirković MM (eds) Global catastrophic risks. Oxford University Press, New York, pp 91–119
Yudkowsky E (2011) Complex value systems are required to realize valuable futures. In: Schmidhuber J, Thórisson KR, Looks M (eds) Artificial general intelligence: 4th international conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings, pp 388–393. https://intelligence.org/files/ComplexValues.pdf. Accessed 25 Apr 2019
Yudkowsky E (2015) Harry Potter and the methods of rationality. http://www.hpmor.com. Accessed 25 Apr 2019