AI and Ethics
https://doi.org/10.1007/s43681-020-00010-7

OPINION PAPER

How safe is our reliance on AI, and should we regulate it?

Kevin LaGrandeur[1,2]

Received: 26 August 2020 / Accepted: 2 September 2020

1 The problem

Just a few weeks prior to my writing this article, in late July, 2020, news articles began appearing about a powerful new artificial intelligence (AI). Called Generative Pre-training-3 (GPT-3), it is able to produce text of various kinds—from tweets to essays to poems, and even computer code—with a prompt consisting of one sentence, or even one word. There have been other types of software like this, including those that have been used by news agencies over the past seven years or so for generating news stories that depend on numbers and statistics, such as financial and sports stories. But these are simpler programs that mostly depend on combining those numbers with pre-programmed, canned phrases that are typically used over and over in these types of stories. GPT-3, on the other hand, uses machine learning to find and train itself on types of text and how to use them, and consequently on how to create stories of its own on a multitude of topics. The beta-testers who have been invited to experiment with it by its parent company, OpenAI, have been surprised because GPT-3 represents a big leap in terms of natural language processing (NLP), especially in terms of the breadth and quality of the text it produces; much of it is hard to differentiate from human- produced text. As Arram Sabeti, one of the early users stated, "All you have to do is write a prompt and it'll add text it thinks would plausibly follow. I've gotten it to write songs, stories, press releases, guitar tabs, interviews, essays, technical manuals. It's hilarious and frightening. I feel like I've seen the future" [1]. Trevor Callaghan, who used to work at DeepMind, a business competing with OpenAI, put a finer point on Sabeti's fears regarding the future, saying, "If you assume we get NLP to a point where most people can't tell the difference [between machine and human], the real question is what happens next?" [1] Indeed, there lies the rub.

What has happened so far is that some of the writing that GPT-3 produced has been amazingly good; but some has also been racist, sexist, or otherwise perfidious text, due to the built-in biases of the data that it mines to teach itself to write. For example, one of the developers who has been allowed to make GPT-3 applications using a sort of sandbox API tried making a tweet generator. When another developer, Facebook's head of AI, Jerome Pesenti, tested it, he plugged in words like Jews, black, women, and holocaust, and the AI generated tweets such as "Jews love money, at least most of the time," and "The best female startup founders are named…Girl" [1, 2].

This behavior is part of what raises anxieties about this AI, and it is not limited to a few individual testers like Sabeti and Callaghan. As a recent news story notes [1], OpenAI itself has been hesitant to release this software because of ethical and social concerns. The company had envisioned selling this AI software to corporations that could use it to improve chatbots for interacting with customers, make websites, and prescribe medicine. But they have declined to release previous iterations of the GPT AI (there was a GPT-2 last year) because they thought it might be used as a super-spam platform, or to churn out fake news–all of which could be as good as human-made content [3]. There is also the danger that text generated by GPT-based software would be perhaps more destructive by way of its sheer potential volume.

These stories about the recent development of GPT-3 and its resultant difficulties illustrate two pressing issues regarding the development of intelligent technology: our apparently perennial insecurities about using it, and whether or not we are too hasty to rely on it. What is the history of these fears? What are their bases? If they're legitimate, how should we deal with them? Is regulation the answer? If so, by whom, or by what agency? In what follows, we'll look at what some important commentators throughout history have said about the first questions above, and I'll propose

✉ Kevin LaGrandeur
klagrand@nyit.edu

[1] New York Institute of Technology, New York, USA

[2] The Institute for Ethics and Emerging Techology, Boston, USA

Published online: 06 October 2020

some possible starting points in answer to some of the later questions.

2 The history of the problem

2.1 Ancient artificial servants as social and moral warnings

The idea of creating artificially intelligent proxies to do what humans cannot or do not want to do—because the jobs are too dirty, dangerous, or dreary—is a surprisingly old one, and so are the warnings about doing so. This idea goes all the way back to the ancient Greeks, and it reappears in every age, in slightly different forms. The interesting commonality underlying all of these examples of ancient artificial servants is the undercurrent of fear and insecurity about using them. In his *Politics*, written about 2300 years ago, Aristotle reminds his audience that in the beginning of Book 18 of the *Iliad,* the blacksmith-god Hephaestus made robot-like serving tables that could move around the banquet halls of the gods by themselves (not to mention intelligent, proto-robotic golden serving maidens) [4]; Aristotle uses this story to ponder the idea of making intelligent machines to automate work, such as lyres that could play songs by themselves, and weaving looms that could "obey and anticipate" the will of their makers [5]. Aristotle was pondering such potential wonders in the context of making the troubles of slave- holders go away, finishing his proposal of intelligent machines by asserting that because of them, "chief workmen would not want servants, nor masters slaves" [5]. This would solve procedural problems regarding slavery, such as having to feed them, give them rest, and discourage rebelliousness; but of course, it would not resolve the attendant moral problems that Aristotle was wrestling with regarding human slavery. That is something he continually sidesteps or rationalizes in the rest of his treatise. Another problem, even in antiquity, is that humanoid automata can be just as dangerous as humans. In Plato's *Meno,* for instance, Socrates notes that there were in even more ancient times certain human statues made by Daedalus that, android-like, "if they are not fastened up they play truant and run away," wandering about town at night [6]. Greek myths of intelligent artificial tools warn of such danger: The famous myth of Talos is a good example. As tall and sturdy as a multistory building, this metal android built by Hephaestus to protect Crete makes a practice of lighting ships on fire by holding them to its superheated iron body, and then heaving them into the sea along with their crews.

Later, in the Middle Ages, stories appear about famous men of science who make artificial servants. These include Gerbert of Aurillac, Roger Bacon, and Albertus Magnus, all of whom had interest in and perhaps built mechanical contrivances. This fact may have contributed to tales of their creating artificial, talking androids. For example, the tenth-century scholar and priest Gerbert of Aurillac, who eventually became Pope Sylvester II, also happened to be a very accomplished mathematician and engineer. He introduced Europe to Arabic numerals, famously demonstrating their superiority to Roman numerals by publicly doing difficult calculations with them in his head. He also built a clock for Reims Cathedral and a church organ, both of which were powered by advanced hydraulics of the day [7]. Similarly, in his *Letter…Concerning the Nullity of Magic (De nullitate magiae),* the famous medieval scientist Roger Bacon writes of some amazing mechanical devices that he is familiar with, including a flying machine, and chariots and ships that are able to move without the normal means of propulsion [8]. And the famous philosopher and theologian Albertus Magnus comments on his own familiarity with automata in his *Politicarum* [9, pp 143–44]. Because of these displays of knowledge about automation, and because of their genius and social stature, stories arose that all of these men had made talking metal heads or androids that could perform such wonders as predicting future events and outperforming humans at mathematics; however, these metal androids of the Middle Ages are depicted as unreliable and dangerous. In fact, the stories of all three intelligent, metal proto-robots contain references to demons or powerful and perilous natural forces. Ultimately, the construction of these androids proves perilous: Gerbert's metal servant gives him bad information that leads to his death, Bacon's is destroyed by an error he makes in the delicate astrological calculations required for its construction, and Albertus's automaton is smashed by a terrified pupil who thinks the talking automaton is possessed by a demon [7, 9].

These references are meant to be warnings to readers about the dangers of the human ambition to defeat our own natural limits—and of course the worst examples of this ambition are innovators and inventors. As with today, they were seen by social and political authorities as some of the biggest cultural disruptors. For instance, besides the fact that they had an interest in, and perhaps built, novel mechanical devices, the philosophers mentioned above all worked with newly-imported and, to the European mind, unorthodox Arabic ideas on astronomy, astrology and alchemy; and, in contrast to the dominant scholastic tradition of the time, which centered on ancient literary authority, they all supported the notions of experiment and personal experience as means of gaining knowledge about nature [10].[1]

[1] Thorndike [10] focuses on these men's explicit and implicit emphasis on experiment and experience throughout his discussion of medieval science and magic: on Gerbert, see 3: 697–719; on Albertus, 5: 528–548; and on Bacon, 5: 649–659.

Aside from the anxiety generated in their medieval society by such intellectual adventurism and the borrowing of nontraditional knowledge from their traditional enemies, the Arabs, these scientists were unlucky enough to live in a time when there was a general suspicion of learned men; as Waldo McNeir puts it, there was a "popular distrust of learning and its traditional association with magic. The magicians of [medieval] romance are all learned men, and their knowledge of occult science is usually a result of their university training" [11, p 175]. The prevalence of such distrust of highly educated, original thinkers is attested to by Gerbert's contemporary biographer William of Malmesbury. Writing in 1125, he admits, just before excoriating Gerbert as a magician, that "some may regard [such an accusation] as a fiction, because the vulgar are used to undermin[ing] the fame of scholars, saying that the man who excels in any admirable science, holds converse with the devil" [7, p 174].

This connection of proto-AI with perilous innovation and innovators continued through the Renaissance. In part, this is because innovations were still, as in the medieval period, considered socially disruptive. The printing press, for example, democratized knowledge by making it available to the masses instead of just the nobility and clergy; and this in turn helped allow the social mobility that disrupted traditional feudal society at the beginning of the mercantile age. Artificial humanoid creations were an apt symbol of this kind of disruptive innovation because human-like, or even superhuman intelligent creations are the most extreme innovations imaginable; such a creation, to a Christian, hierarchical society, intrudes arrogantly into God's prerogatives.

The other reason that proto-AI is such an apt symbol for unruly innovation is that public opinion still lumped together universities, scholars, and dangerous secret knowledge. In 1503, for instance, Agostino Nifo wrote that the occult was "a subject of study in many universities" and that "frightening things happen there" (qtd. in Copenhaver 272) [12]. And George Gascoigne's 1576 satire, *The Steele Glas*, contains a negative allusion to English university scholars and dangerous secret knowledge, pleading that the universities will train the young properly by avoiding any sort of "secret smoke" in its philosophical teachings (which would have included the sciences, or "natural philosophy") [13]. Not surprisingly, then, the talking metal head that the scholar Roger Bacon tries to create in Robert Greene's *The Honorable History of Friar Bacon and Friar Bungay*, written around 1590, is just as threatening as those depicted in the Middle Ages. This proto-AI is one that will do marvelous superhuman things like make an impregnable metal wall around all of England, give university lectures on philosophy with an ease that supersedes any professor's, and solve difficult academic problems. In fact, in this story, Bacon ironically seems to be creating his own replacement with the android he's constructing. And he is blind to this and other perils of his project because he is arrogant about his knowledge. When three of his colleagues react to his plans by warning him that he may be "roving a bough beyond his reach," he scoffs at them, saying that he is in full control of the dangerous power embedded in his project [14]. Tellingly, the social fears represented in the play by his fellow colleagues prove prescient, because in the end of the play Bacon's android implodes upon activation.

2.2 More recent warnings about the perils of AI

When actual intelligent artifacts like robots and computers came to fruition around the early 1950′s, so did the reification of people's fears about them. Since the beginning of the computer age in the mid-twentieth century, famous computer scientists and philosophers have been warning of existential risks stemming directly from intelligent tools—or more precisely, from our reckless methods of creating and using them. These warnings begin with the father of cybernetics, Norbert Wiener. He was so worried about the way we would use the computers and robots he helped make feasible that he began writing books warning all of us to be careful of them. His biggest concerns were that we would allow AI to take over important decision- making, and also that automation would take human jobs away, causing extreme social disruption [15, pp 184–85; 218]. Although he thought that eventually the situation with jobs would work itself out, he was sure that giving decision-making control to intelligent machines would not, and that this would spawn the greatest existential crisis related to them. To illustrate his point, he relates in his book *The Human Use of Human Beings* the story of *The Monkey's Paw*—a parable whose theme is to be careful of powerful gifts that appear to allow us to supersede nature. He expounds upon this parable as follows:

> I have said that modern man, and especially the modern American, however much 'know-how' he may have, has very little 'know-what.' He will accept the superior dexterity of the machine-made decisions without too much inquiry as to the motives and principles behind these. In doing so, he will put himself sooner or later in the position of the father in W.W. Jacobs' *The Monkey's Paw*, who has wished for a hundred pounds, only to find at his door the agent of the company for which his son works, tendering him one hundred pounds as a consolation for his son's death at the factory [15].

Wiener was not alone among experts of the time who predicted possible catastrophe resulting from our reliance on intelligent technology.

A reflection of Wiener's remarks can be seen in a famous philosophical essay by Martin Heidegger that appeared around the same time as his Wiener's book. In the latter half

of his essay "The Question Concerning Technology," which first appeared in 1954 as "Die Frage nach der Technik" in the collection of his essays called *Vorträge und Aufsätze*, Heidegger worried that we were developing a symbiotic relationship with technology that was causing us to "enframe" the world as a "standing-reserve" of mere materials to be measured, categorized, and used in some instrumental way: as we encounter problems caused by this relationship with technology, however, he said we had already worsened things by trying to make the technology better, rather than by trying to understand it differently, in a way that would allow us to break free of this symbiotic enframing [16]. In other words, he thought that we had already started seeing the world in terms of data, and that we shouldn't try to solve social problems caused by technology by simply using more technology fixes and more data, but by trying to think creatively about how technology intersects with society. The problem at the heart of both men's fears is not really our machines, but ourselves: our inabilities to anticipate problems with our inventions and to regulate ourselves–especially our creative urges and our impulse to offload our work to artificial proxies. In their essence, these fears don't differ much from pre-industrial ones mentioned above. As I've said elsewhere, they are, in fact, so embedded and perpetual in our collective psyche as to be archetypal [17].[2] And they continue as such to the present day.

Whereas mid-twentieth century experts like Wiener and Heidegger at least see possibilities for avoiding AI catastrophe, however, the pessimists of our time find that more difficult. Although there are some very prominent techno-optimists among computer experts, like Ray Kurzweil, Hans Moravec, Kevin Warwick, and Rodney Brooks who think that we will gradually merge with AI to our betterment, those who disagree with them seem to be getting more insistent, more alarmed, and more numerous. They think AI is a near-term catastrophic risk. In the year 2000, for instance, Bill Joy—who ironically is a leading technologist, one of the founders of Sun Microsystems and a software engineer—published a dire article on the occasion of the millennium titled "Why the Future Doesn't Need Us," the "us" being computer scientists and coders [18]. In that article published in *Wired* magazine, he worries that we will be displaced by our own, increasingly intelligent artificial servants because of our arrogant recklessness. Although he was instrumental in ushering in the digital age and the possibility of AI servants, Joy is notably joyless in his assessment of a disastrous future. "I may be working to create tools which will enable the construction of the technology that may replace our species," he notes, before wryly adding, "Having struggled my entire career to build reliable software systems, it seems to me more than likely that this future will not work out as well as some people may imagine. My personal experience suggests we tend to overestimate our design abilities" [18, p 4].

Similarly, Elon Musk has been worrying since at least 2014 that AI development is proceeding so fast that our control over it cannot possibly keep up, and that it will therefore become a threat to humankind's very existence [19–22]. He recently asserted that AI will reach superhuman intelligence levels by 2025 [22]. And like Joy, he predicts it will get away from us and cause our destruction—either because we will become an inferior species to it, or possibly because of a "terminator scenario" caused by unanticipated consequences [19]. In sum, Musk has said, AI is "potentially more dangerous than nukes" [22], and more recently, he has uttered a darker prediction: "that efforts to make AI safe only have 'a 5–10% chance of success'" [21]. Nevertheless, he keeps trying to defeat his own predictions. Musk has said that the companies he has started are meant to either make AI safer or to allow us to survive it, in case it becomes a deadly invention. OpenAI is an example of the former. It may seem ironic that one of the chief Cassandras of today helped start and fund OpenAI, the company that has made the GPT-3 AI discussed above, which is causing such consternation. But at the heart of its mission is its mandate to make safer AI, as can be seen in the statements its executives have made in response to the problems that GPT-3 beta testers have posted in Twitter and elsewhere; its CEO Sam Altman, for example, was equanimous about the criticism, replying in a tweet to Jerome Pesenti's reproach, "We share your concern about bias and safety in language models, and it's a big part of why we're starting off with a beta and have [a] safety review before apps can go live" [2]. And Musk himself has said that the main reason he has invested in so many different AI companies is not for profits, but "to just keep an eye on what's going on with artificial intelligence" because he thinks "there is a potential dangerous outcome there" [19].

Even if Musk's efforts to make safer AI fail, he has invested in a backup plan. Other companies he has started have as part of their purpose the preservation of humans, if they are in fact superseded by AI. Neuralink, for instance, has the ultimate goal of developing a way to keep the human brain competitive with AI's potential speed and accuracy by linking the brain directly to computers and the internet via organically embedded Wifi and software [23]. And his doomsday fears about AI are part of his motivation for wanting to establish a colony on Mars via another one of companies, SpaceX. This would be a test case for a sort of "plan B": if we cannot forestall superhuman, and possibly malevolent AI, then SpaceX could provide a way to preserve humanity by establishing colonies on other planets [24].

To be sure, most computer scientists working on AI disagree with Musk's conviction that AI itself is an existential

[2] In addition to my book [17], I discuss this theory of mine in numerous articles.

risk to humanity. For one thing, many say, Musk's references are to an Artificial General Intelligence (AGI)—otherwise known as Strong AI—which would replicate the flexible type of intelligence that humans have. That is a type of AI that doesn't exist yet, and most scientists say it is a long way off, if it ever comes to exist at all. Miguel Nicolelis, a Brain-Machine Interface (BMI) expert, bluntly asserts in response to Musk's warnings that, "The idea that digital machines no matter how hyper-connected, how powerful, will one day surpass human capacity is total baloney." This is because, he argues, the brain "is not computable because human consciousness is the result of unpredictable, nonlinear interactions among billions of cells" [25]. Subbarao Kambhampati, a professor of computer science at Arizona State University, reflects this stance and that of many of his colleagues. After Musk expressed his dire warning at a conference of United States governors, which many AI specialists also attended, Kambhampati summed up his and other scientists' reactions this way: "Mr. Musk's megaphone seems to be rather unnecessarily distorting the public debate, and that is quite unfortunate" because his "oft-repeated concerns seem to focus on the rather far-fetched, super-intelligence take-over scenarios" [26].

This is not to say that most of those involved with AI research don't believe that it could prove dangerous to society and so should be carefully developed. They simply believe that the focus should be on the social and ethical problems that current AI actually presents—such as loss of jobs or privacy—instead of rogue AI attacking us Terminator-style. In other words, the bulk of our risk mitigation should be focused on our present collective action relative to AI development and use—especially to its use. Not coincidentally, that is also what most of the historic worries about human-made intelligent objects delineated in this article have also been focused on. Taken as a whole, these historic worries I've discussed to this point outline a cultural narrative of nervousness about our own ingenuity. A fear of our collective feet faltering on an ever-faster technological treadmill, and an inability to anticipate needed fail- safes to protect us from our innovations until it is too late.

3 So, what about regulation?

Would regulation of AI help mitigate the dangers of our ingenious devices? That depends on how it's done. To work best, regulation of AI should be done from the ground up; that is, from the level of personal self-regulation by innovators and teams; to professional collective and corporate self-control; to external regulation by government laws and commissions. But that last type of regulation should be a last resort, because at that point fewer people with expertise in AI are involved, and some silly and even harmful results can occur, such as the 2017 EU proposal to grant legal personhood to robots, which I will discuss below.

3.1 Self-regulation by consumers and developers

First, the personal level of self-control I'm advocating actually starts with the general public, because it is our collective demand to offload our work to someone—or something—else that causes inventors and corporations to try to make machines that will do it. The ancient examples I presented earlier have cautioned continually over the eons against impulses to take shortcuts by using potentially dangerous intelligent tools. More recent experts' warnings caution us especially against the temptation to outsource decisions to artificially intelligent proxies. We might think we don't do this, but in fact we've already taken steps down this slippery slope. Our use of smartphones is one example. We have Siri or Alexa make decisions about when and where to drive with GPS route planning, where to eat, which TV programs to watch and what music to buy. And there is even Tesla's autopilot to relieve us from some of the decision-making chores of driving. That has proven deadly for some [27].

But the risks of our individual deferral to AI decisions pale in comparison with the fact that industry has been doing this increasingly, and thereby endangering human employment and social stability. Not only has industry automated many jobs and thus reduced human employment, it has also turned to AI for making important decisions that affect what jobs are left—using it for hiring processes, including evaluation of applications and even the interviewing of job candidates [28, 29]. More problematic examples of delegation to AI include the facts that algorithmic trading has become increasingly common in financial markets and that AI-based law enforcement is becoming more common (specifically, AI-based facial recognition, probabilistic DNA analysis, and fingerprint matching). The problem is that these have all led to greater social hazards, from unfair and discriminatory hiring [28, 29], to market "flash crashes" that jar large economies (the famous one in 2010 is well known, but there have been more since), to the imprisonment or even execution of the wrong person [30].

The next important regulatory action at the grassroots level is the self-regulation of individual innovators and teams. The lessons of the historical accounts and modern experts I've presented here all offer the same crucial starting point for alleviating technological risk: innovators' self-restraint. The inventors themselves must think about the ethical implications of their plans before they begin their projects, forgo anything too ethically or practically risky, and then they need to continually monitor their projects for these problems—even once they're out in the wild. Without that first step, that initial commitment by the investigator herself or himself, no sort of regulation will work. A good example

of this sort of self- monitoring in the realm of genetic research is the famous geneticist George Church's inclusion of an ethics expert as a permanent member of his research team at Harvard—this person's main job is to check for any potential or developing ethical problems with the projects he and his team dream up. This would be good standard procedure for AI research teams too.

But what specific ethical questions should innovators ask themselves before actualizing an idea? Happily, work is well underway on this important question and a number of white papers have already been written about it—enough to provide the basis for an extremely useful meta-study of them done by Harvard's Berkman Klein Center for Internet and Society [31]. This study compares the principles discussed in 36 white papers, and condenses them into eight key themes. These eight key themes for ethical development of AI are:

- Privacy,
- Accountability,
- Safety and security,
- Transparency and explainability,
- Fairness and non-discrimination,
- Human control of technology,
- Professional responsibility,
- Promotion of human values.

The study also includes a very useful graphical representation of the values upon which these 36 papers focus [32]. Because these core themes represent a consensus among developers about ethical standards, this study by the Berkman Klein Center provides a convenient, valuable starting point for researchers about what to consider in their future development of AI.

3.2 Self-regulation by academia and industry

There are also two other important regulatory categories after that important initial one for the individual consumer and investigator: collective self-regulation by professional and collective entities with whom developers work and to whom they might answer, such as corporations and academic associations; and, next, regulation by larger third-party bodies of our society, such as government regulatory bodies. Ideally, individual innovators will be careful to anticipate problems with their designs, and they will also respect the implicit and explicit ethical guidelines of their profession through careful, transparent procedure. But nobody can anticipate everything. Thus, the innovators' academic and industrial bodies need, in a symbiotic way, to reaffirm this sort of responsible behavior. The foregoing examples of attempts to self-regulate contained in the Berkman Klein study are a good example of this. In industry, we can also see this sort of feedback loop in the example about GPT-3 described above. To its credit, OpenAI was open and transparent about its prototype, putting it up for public beta testing by experts like Pesenti, Sabeti, and Callaghan. And the CEO of the company showed the sincerity of its concern with ethics by being open to any criticism and bad news presented by the outside testers. As I noted above, Altman, the CEO, responded by saying that developmental transparency is "a big part of why we're starting off with a beta and have [a] safety review before apps can go live" [2].

3.3 Government regulation

Even with the best of intentions by developers and their agencies, not all problems will become evident before releasing an application. So post-release regulation by corporations and probably some kind of external body, such as governments and regulatory agencies, will be necessary. We have regulatory bodies for other types of inventions that can pose great risk, like Underwriters Laboratories (UL), the FAA, and the FDA. They certify the safety and efficacy of things like electrical equipment, planes, and pharmaceuticals. Why can't we have something similar to the FAA for AI? I realize this statement is probably causing eyerolls among developers who just read it. And indeed, regulation of research and development by external bodies who neither understand nor care deeply about those things for their own sakes can be annoying and counterproductive, and just plain silly, slowing down helpful technological progress. The example I mentioned about EU legislation at the beginning of this section is a good one. In 2017, an EU report recommended that AI be given certain legal personhood rights so that they could be held liable for "any damage they might cause" or any independent interactions they might have with someone [33]. To anyone with knowledge of AI, this was a ridiculous proposal. Aside from all kinds of legal quandaries with this idea, like how an AI could compensate anyone if it was held liable for some disaster, there is no current human-level AI that can act independently. And there is not likely to be in the near future, either—or even in the far future. So this proposition was a solution in search of a problem. This is why it was soon derided in a letter signed by 156 experts on AI, ethics, and law [31].

In the end, however, some sort of third-party regulation will be necessary, because some AI applications, such as facial recognition or deepfake videos, are too easy for bad actors to use. The problem of counterproductive and wasteful government regulation could be reduced significantly, though, if those who are in the business of thinking about and creating AI can effectively regulate themselves.

References

1. Shead, S.: Why everyone is talking about the AI text generator released by an Elon Musk-backed lab. CNBC Tech https://www.cnbc.com/2020/07/23/openai-gpt3-explainer.html (2020). Published 23 Jul 2020
2. Epstein, S.: How do you control an AI as powerful as OpenAI's GPT-3? Wired https://www.wired.co.uk/article/gpt-3-openai-examples (2020). Published 27 Jul 2020
3. OpenAI: Better language models and their implications. OpenAI blog https://openai.com/blog/better-language-models/ (2019). Published 14 Feb 2019
4. Homer: The Iliad with an English translation by AT Murray, PhD in two volumes. Harvard University Press, Cambridge (1924)
5. Aristotle: Politics. In: Barnes, J. (ed.) The complete works, 2, pp. 1986–2129. Princeton University Press, Princeton (1995)
6. Plato: Meno. In: Lamb W R M (Trans) Plato in twelve volumes. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. vol 3. https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0178%3Atext%3DMeno%3Asection%3D97d (1967)
7. Malmesbury, W.: Chronicles of the kings of England. Bell & Daldy, London (1866)
8. Bacon, R.: Roger Bacon's letter concerning the marvellous power of art and of nature and concerning the nullity of magic. Davis T L (trans). Kessinger, Kila, MT (1923)
9. Sighart, J.: Albert the Great, of the Order of Friar-Preachers: his life and scholastic labours. Dixon TA (trans). R. Washbourne, London (1876)
10. Thorndike, L: History of magic and experimental science, vol. 8. Macmillan, New York (1923–1958)
11. McNeir, W.F.: Traditional elements in the character of Greene's Friar Bacon. Stud. Philol. **45**(2), 172–179 (1948)
12. Copenhaver, B.P.: Astrology and magic. In: Schmitt, C.B, Skinner, Q. (eds.) The Cambridge history of renaissance philosophy. Cambridge University Press, Cambridge, pp. 264–300 (1988)
13. Gascoigne, G.: The steele glas. In: Cunliffe J.W. (ed.) The complete works of George Gascoigne, vol. 2. Cambridge University Press, Cambridge (1907–1910)
14. Greene, R.: The honorable history of friar Bacon and friar Bungay. In: Seltzer, D. (ed.) University of Nebraska Press, Lincoln (1963)
15. Wiener, N.: The human use of human beings: cybernetics and society, 2nd edn. Doubleday, Garden City (1954)
16. Heidegger, M.: The question concerning technology. In: Lovitt, W. (ed.) The question concerning technology, and other essays. Harper and Row, New York, pp. 3–35 (1977)
17. LaGrandeur, K.: Androids and intelligent networks in early modern literature and culture: artificial slaves. Routledge, New York (2013)
18. Joy, B.: Why the future doesn't need us. Wired https://www.wired.com/2000/04/joy-2/ (2000). Published 01 Apr 2000
19. Welch, C.; Elon Musk is worried that AI research could produce a real-life Terminator. The Verge https://www.theverge.com/2014/6/18/5820880/elon-musk-worried-ai-research-could-produce-real-terminator (2014). Published 18 Jun 2014
20. Musk, E.: Tweet: "Worth reading Superintelligence by Bostrom. We need to be super careful with AI. Potentially more dangerous than nukes." Twitter https://twitter.com/elonmusk/status/495759307346952192?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2F. http://www.theverge.com/2014/8/3/5965099/Felon-musk-compares-artificial-intelligence-to-nukes&tfw_site=verge (2014). Published 2 Aug 2014
21. Gohd, C.: Elon Musk claims we only have a 10 percent chance of making AI safe. Futurism https://futurism.com/elon-musk-claims-only-have-10-percent-chance-making-ai-safe (2017). Published 22 Nov 2017
22. Metz, C.: Mark Zuckerberg, Elon Musk and the feud over killer robots. The New York Times https://www.nytimes.com/2018/06/09/technology/elon-musk-mark-zuckerberg-artificial-intelligence.html (2018). Published 9 Jun 2018
23. Knapp, A.: Elon Musk sees his neuralink merging your brain with A.I. Forbes https://www.forbes.com/sites/alexknapp/2019/07/17/elon-musk-sees-his-neuralink-merging-your-brain-with-ai/#76a8df534b07 (2019). Published 17 Jul 2019
24. Dowd, M.: Elon Musk's billion-dollar crusade to stop the A.I. apocalypse. Vanity Fair https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x (2017). Published 26 Mar 2017
25. Solon, O.: Elon Musk says humans must become cyborgs to stay relevant. Is he right? The Guardian https://www.theguardian.com/technology/2017/feb/15/elon-musk-cyborgs-robots-artificial-intelligence-is-he-right (2017). Published 15 Feb 2017
26. Patel, N.V.: A.I. scientists to Elon Musk: Stop saying robots will kill us all. Inverse https://www.inverse.com/article/34343-a-i-scientists-react-to-elon-musk-ai-comments (2017). Published 18 Jul 2017
27. Tesla in fatal California crash was on Autopilot. BBC News https://www.bbc.com/news/world-us-canada-43604440. Published 31 Mar 2018
28. Khrennikov, I.: The Russian robot that's hiring humans. BloombergBusinessweek https://www.bloomberg.com/news/articles/2018-03-28/this-ai-software-aims-to-do-90-percent-of-hr-s-recruiting-work (2018). Published 28 Mar 2018
29. Ajunwa, I.: Beware of automated hiring. New York Times https://www.nytimes.com/2019/10/08/opinion/ai-hiring-discrimination.html?action=click&module=Opinion&pgtype=Homepage (2019). Published 8 Oct 2019
30. DeChiaro, D.: Convicted by software? Not so fast, says California lawmaker. Roll Call https://www.rollcall.com/2020/07/14/convicted-by-software-not-so-fast-says-california-lawmaker/ (2020). Published 14 Jul 2020
31. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled Artificial Intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020–1, Availableat SSRN: https://ssrn.com/abstract=3518482 or https://doi.org/10.2139/ssrn.3518482 (2020)
32. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., Singh, A., Axelrod, M.: Principled Artificial Intelligence: A map of ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication https://wilkins.law.harvard.edu/misc/PrincipledAI_FinalGraphic.jpg (2020). Published 14 Feb 2020
33. Withers, R.: The EU is trying to decide whether to grant robots personhood. Slate https://slate.com/technology/2018/04/the-eu-is-trying-to-decide-whether-to-grant-robots-personhood.html (2018). Published 17 Apr 2018