AI Anxiety

Deborah G. Johnson

University of Virginia

Mario Verdicchio

University of Bergamo

Authors Note:

Deborah G. Johnson, Department of Engineering & Society, University of Virginia.

Thornton Hall, PO Box 400744, Charlottesville, VA 22904.

Phone: +1 434-924-7751

Email: dgj7p@eservices.virginia.edu

Mario Verdicchio, Department of Management, Information and Production Engineering,

University of Bergamo.

viale Marconi 5, 24044, Dalmine (BG), Italy.

Phone: +39 035-2052-358

Email: mario.verdicchio@unibg.it

Contact Author: Deborah G. Johnson

Abstract

Recently a number of well-known public figures have expressed concern about the future development of Artificial Intelligence (AI), by noting that AI could get out of control and affect human beings and society in disastrous ways. Many of these cautionary notes are alarmist and unrealistic, and while there has been some pushback on these concerns, the deep flaws in the thinking that leads to them have not been called out. Much of the fear and trepidation is based on misunderstanding and confusion about what AI is and can ever be. In this work we identify three factors that contribute to this 'AI anxiety': an exclusive focus on AI programs that leaves humans out of the picture, confusion about autonomy in computational entities and in humans, and an inaccurate conception of technological development. With this analysis we argue that there are good reasons for anxiety about AI but not for the reasons typically given by AI alarmists.

Keywords: Artificial Intelligence, Autonomy, Sociotechnical Systems, Future

AI Anxiety

Recently a number of well-known public figures have expressed concern about the future development of Artificial Intelligence (AI). Bill Gates, Elon Musk, Stephen Hawking and others have noted that AI could get out of control and affect human beings and society in disastrous ways (FLI, 2015). Although these expressions of concern are to be lauded in principle because they draw attention to the social implications of rapidly developing, enormously powerful computational capacities, many of these cautionary notes are alarmist and unrealistic. While there has been some pushback on these concerns, the deep flaws in the thinking that leads to them have not been called out. For example, Roger Schank recently noted in his blog post "Hawking is afraid of A.I. without having a clue what A.I. is" (Schank, 2016) that Hawking is afraid of a technology that behaves like human beings, a technology that reasons by means of something like human mental processes. This only begins to get at the confusion that surrounds AI.

Fear and trepidation about AI is also expressed in futuristic scenarios put forward in science fiction films and novels. Here we see a number of scenarios played out: humans wrestling with the status of robots (e.g., Spielberg's *AI* and Jonze's *Her*); robots with moral sensibilities choosing the lesser of two evils (e.g. Proyas's "I, Robot"); and intelligent robots yearning for freedom (e.g. Garland's "Ex Machina").

Although such futuristic scenarios are generally put forward as science fiction, some literature argues that the AI/robot takeover is a real possibility, if not an inevitable end-state of what is happening now. For example, in Nick Bostrom's "Superintelligence" (Bostrom, 2014), we are told that in the future, AI will not just improve beyond its current capacity, but will improve its capacity to improve itself. After a 'covert preparation phase', in which the ever improving, 'strongly superintelligent' technology will secretly devise a plan to achieve its long-term goals, it will enter a final phase of 'overt implementation', in which the AI will be sufficiently strong to obviate the need for secrecy and will start a 'strike' against humanity. According to Bostrom, at this point the AI will be an 'autonomous' technology that has gone out of human control, acquiring new goals without any direction from its original designers and procuring the necessary resources to achieve them. Given that humans will have no say in the superintelligent AI's decision making, humans will be irrelevant, if not enslaved or extinct.

The simple phrase ‘AI anxiety’ can be used to refer to the fear and trepidation being expressed about out of control AI. AI anxiety actually has a long history going back to the first modern computers when many feared that computers threatened the very idea of what it means to be human. Human beings are thought to be unique because of their capacity to think, and if computers also can think, then fundamental notions about what it means to be human are undercut.

The challenge of distinguishing humans and machines has fascinated and troubled scholars for a long time. Some try to draw a line in terms of activities that only humans can do, i.e., activities that are out of the reach of a machine. What has become known as the ‘Lovelace objection’ (“computers can do many things, but not X”) is a constant presence in debates about AI. For example, in Dreyfus’ defense of human superiority in his 1978 “What Computers Can’t Do” (Dreyfus, 1978), he argues for the impossibility of modeling human commonsense by means of a finite sequence of symbolic instructions. After the advent in the early 1980s of neural networks, which seemed to avoid many of the problems related to symbolic approaches in traditional AI, Dreyfus extended his critique in his 1992 “What Computers Still Can’t Do” (Dreyfus, 1992). Fascination with the line between humans and AI machines can be seen more recently in IBM’s creation of Deep Blue, which was pitted against world champion chess player Gary Kasparov. Google’s AlphaGo program is yet another attempt to pit human intelligence against AI.

Stepping back, one can’t help but wonder why there is such anxiety about AI and its future development. We believe that there are good reasons to be concerned about the directions of AI development but we also believe that much of AI anxiety is misdirected. Much of the fear and trepidation is based on misunderstanding and confusion about what AI is and can ever be.

Sociotechnical Blindness

Expressions of AI anxiety tend to abstract AI out of the context of its existence and use, ignoring the human beings and the social institutions and social arrangements that give AI its capacity to do or mean anything. On the one hand, AI can be understood to be lines of code that perform computational operations. As such, AI does nothing that affects the world. To have effects in the world, AI programs have to be put into physical computers that perform operations that have

meaning for humans. Consider a furnace that is turned on by signals triggered by a digital output, an airplane that changes direction when software operates, a bid for a particular stock by a software agent for stock market trading. The behavior of programs has significance only as part of technological systems that serve human purposes.

AI programs can be distinguished from AI *sociotechnical systems*. Programs are, simply put, lines of code. AI systems, instead, consist of those lines of code together with the context in which the code operates. It's the difference between a trading program and a stock exchange in which AI programs are used. In and of itself, a trading program is an artifact but when the program is embedded in a stock exchange, it operates together with human behavior. This is true both in the sense that people have to put the program in the system and in the sense that people have to recognize what the program does as meaningful. While AI software, such as trading software, is a computational artifact, AI systems are sociotechnical.

AI anxiety generally results from an exclusive focus on AI programs leaving out of the picture the human beings and human behavior that create, deploy, maintain, and assign meaning to AI program operations. We refer to this blindness to the human contexts in which programs are used as *sociotechnical blindness*. Those who have sociotechnical blindness fail to recognize that AI is a system and always and only operates in combination with people and social institutions.

Thus, futuristic scenarios like Bostrom's are misleading because they suppose that AI of the future will function without human beings. They suppose that futuristic systems will function without the need for human beings to maintain the physical and social infrastructure in which AI operates. This is like imagining a global stock exchange taken over by trading bots that continue to operate even when there are no human beings interested in making money or no brick and mortar companies behind the stocks.

Sociotechnical blindness has many dangers because it enables unrealistic scenarios in which humans are out of the picture. For AI to develop into any futuristic form, human actors will have to make a myriad of decisions on what research to invest in and what software and hardware to develop. Human actors will also have to decide what contexts to embed AI programs in and what social arrangements to set up to launch, monitor, and maintain the AI computational artifacts. Hence, no future development is inevitable simply from the nature of computation.

Confusion About Autonomy

In addition to sociotechnical blindness, AI anxiety results from confusion about autonomy, that is, from thinking of AI as autonomous but not thinking much about what counts as autonomy. Autonomy in computational entities is very different from autonomy in humans, though in AI anxiety the former is often conflated with the latter. For humans, autonomy refers to the characteristic of having the capacity to make decisions, to choose, and to act. This type of autonomy is tied to ideas about human freedom and has traditionally been used to distinguish humans from other types of living and non-living entities, e.g., animals, rocks. Importantly, this form of autonomy is what makes human beings moral beings. Only beings with autonomy can be expected to conform their behavior to rules and laws. It is this notion of autonomy that comes into play in the fear and concern about 'autonomous' AI. When non-experts hear that machines have autonomy, they attribute to machines something comparable to the autonomy that humans have, something close to the freedom to behave as one chooses.

Autonomy in computers can refer to a number of different aspects of these artifacts. For example, it can refer to creation of data on the basis of parameters inside the machine, as in the case of the computer clock used to generate pseudo-random numbers. It can refer to the fact that a computer has sensors that enable it to acquire and operate on data that were not specified by the programmer in the code. In either of the latter definitions, the program is autonomous in the sense that the programmer cannot know in advance the precise outcomes of the program, as when AlphaGo made moves that surprised even its creators. An AI system might also be thought to have autonomy because it has actuators that allow it to move in space and in response to data from the sensors, as in the case of robots exploring the surface of an asteroid or, more simply, a Roomba cleaning the floors of an apartment.

These kinds of autonomy are computational; they are defined in terms of software and hardware processes or characteristics. Whatever human autonomy is, it is clear that it is not the same as computational autonomy. Of course, computer scientists who embrace computationalism would disagree: according to them, human mental processes are computational and, thus, reproducible, at least in theory, on a computer (Scheutz, 2002). Computationalism has not been entirely

refuted, but neither has a proof of its validity been demonstrated to be correct. In either case, it will take a human act – a human decision – to treat a computational configuration as if it is the same as (or comparable to) a human characteristic such as autonomy.

Of course, we can treat computational and human autonomy as metaphorically the same. We can, in effect, use terms like autonomy and agent as short hand for the complex operations in computers. The public, the media, and others do this and it can be quite useful. Nevertheless, the metaphor can lead to misunderstanding. Referring to what goes on in a computer as autonomy conjures up ideas about an entity that has free will and interests of its own – interests that come into play in decision making about how to behave. This then leads to the idea of programming that is insufficient to control computational entities, insufficient to ensure that they will behave only in specified ways. Fear then arises for entities that will behave in unpredictable ways.

An example of this kind of slip from metaphor to sameness is found in Omohundro's chapter in *Risks of Artificial Intelligence* (Müller, 2015). He develops a futuristic scenario in which a robot behaves like a (possibly sociopathic) person who harms others in the blind pursuit of its own objectives. Omohundro describes, for example, a chess-playing robot and a human trying to unplug it: "*Because nothing in the simple chess utility function gives a negative weight to murder, the seemingly harmless chess robot will become a killer out of the drive for self-protection.* (Müller, 2015, p.15)" The drive for self-protection, a natural characteristic of humans and many other biological entities, is presented by the author as a property of advanced AI artifacts. The drive is then supposed to lead to resource acquisition behavior: "*The chess robot* (...) *would benefit from additional money for buying chess books* (...) *It will therefore develop subgoals to acquire more computational power and money. The seemingly harmless chess goal therefore motivates harmful activities such as breaking into computers and robbing banks.* (Müller, 2015, p.16)". In this way, Omohundro slips from the metaphorical similarities between humans and robots to imagine that a chess-playing robot could get so out of control that it would begin robbing banks and killing people.

He gets caught up in the metaphor as if it were real. Omohundro uses 'self-protection', 'drive' and even 'goal' here as if these terms refer to something in computers that is the same as

something in humans. Whatever 'self-protection', 'drive' and 'goal' might mean for a robot, it is not the same as 'self-protection', 'drive' and 'goal' in a human being. Attention is turned away from the computational nature of the robot and the fact that it has been built and programmed to be as it is. Focusing on the fact that the chess-playing robot was built by humans (even though it may have learned since its original programming), should lead us to wonder: *who put the knife in the hand of the robot?* Metaphorically speaking, we might say that the robot behaves like a human psychopath, but it is a robot, so it is comprised of hardware and software. How did it acquire the hardware essential to harming humans? For a robot to kill humans, it would have to have actuators controlled by software that enabled it to grab and maneuver a knife. Sensors would have to gather data not only on the position of the designated victim and her vital organs, but also on the position of the robot parts themselves. Did the robot come spontaneously into existence this way? Clearly not: it is a technological artifact built by humans with these characteristics or, at least, endowed with the tools needed to acquire such skills.

Technological Development

In addition to sociotechnical blindness and confusion about autonomy, an inaccurate conception of technological development also contributes to AI anxiety. Futuristic AI scenarios jump to the endpoint of a path of technological development without thinking carefully about the steps necessary in order to get to that endpoint. Even if computationalism were correct and the endpoint of superintelligent AI with real intentions were possible, the actual path that AI will take in future development is far from clear. The steps required to reach the endpoint that computationalists imagine would have to involve both incredible technological leaps and many, many decisions by human users and designers.

Whatever endpoint in the future we can imagine – be it one in which superintelligence has taken over or not – the sequence of steps preceding the endpoint will have involved humans at some point. Humans would have to make decisions to design AI artifacts in particular ways and delegate operations to them. Yes, the more developed the AI artifacts become, the more distant

human decision making will be from the execution of operations in the pursuit of a goal. Still, however distant human decisions would have been, they will have to have been part of the process at one point or another.

So, the inaccurate view of technological development is intertwined with sociotechnical blindness. In neglecting the role of humans in technological development, AI futurists end up centering their narratives around superintelligent AI artifacts that evolve into dangerous entities completely disconnected from human activities. Humans as well as artifacts are involved in the steps necessary to get to superintelligent AI or any other endpoint in future technological development.

Conclusion

There are good reasons for anxiety about AI but not for the reasons typically given by AI alarmists. AI programs and software should not be the target of our anxiety. The target of our anxiety should be the people who are investing in AI and making decisions about the design and the embedding of AI software and hardware in human institutions and practices. The target should be those who decide when AI programs and systems have been adequately tested, those who have a responsibility to ensure that AI does not get out of control. Decisions made by people can lead to the adoption of inadequately bounded AI; they can lead to AI systems that are unpredictable and even dangerous. The temptation to be negligent may be great because AI seemingly promises to bring the enormous benefits of automation to the next level. Inadequately bounded AI can go out of human control and although it may be possible to turn the machines off, the consequences of doing so will be greater the more dependent we have become on AI.

The possibility that human decision making will lead to AI out of the control of humans and/or AI that behaves in ways that counter human interests is possible and already quite real for certain systems of today. This points to yet another reason for anxiety about the human users and designers of AI. Humans are not a single, monolithic group and do not always have common interests. AI is likely to be developed to serve the interests of some and not the interests of others. This is another reason for keeping an eye on the human actors in the domain of AI systems. Whose interests will be served by the development of increasingly more sophisticated

AI?

The importance of responsible decision-making about AI today and in the near-term term future cannot be overstated. In politics, the old adage is ‘follow the money’; in AI, the adage should be ‘follow the humans.’ Given that *people* will decide what kind of AI we get in the future, fear and trepidation are justified.

References

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

Dreyfus, H. L. (1978). *What Computer Can't Do*. The MIT Press, Cambridge, USA.

Dreyfus, H. L. (1992). *What Computer Still Can't Do*. The MIT Press, Cambridge, USA.

Future of Life Institute (2015). *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*. Retrieved from http://futureoflife.org/open-letter-autonomous-weapons/.

Müller, V. (Ed). (2015). *Risks of Artificial Intelligence*. CRC Press, Boca Raton, USA, 2015.

Schank, R. (2016). *Hawking Is Afraid of A.I. Without Having a Clue What A.I. Is*. Retrieved from http://www.rogerschank.com/hawking-is-afraid-of-ai-without-having-a-clue-what-ai-is.

Scheutz, M. (Ed.). (2002). *Computationalism: New Directions*. The MIT Press, Cambridge, USA.