ESCP
BUSINESS SCHOOL

IT ALL STARTS HERE

BERLIN | LONDON | MADRID | PARIS | TURIN | WARSAW

B. Business Impact

Which tasks shouldn't we delegate to Artificial Intelligence?

ESCP Impact Paper No.2023-11-EN

Sergio VASQUEZ BRONFMAN
ESCP Business School

ESCP RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ESCP Impact Paper No.2023-11-EN

Which tasks shouldn't we delegate to Artificial Intelligence?

Sergio Vasquez Bronfman*

ESCP Business School

Abstract

Since the 1980s recurrent facts show there is a need for ethical thinking on artificial intelligence (AI). Research and institutional policies in this field followed two main directions: the ethics that we should "inject" into AI programs, and the ethics of the use of AI. We argue that teaching algorithms to distinguish right from wrong is too complex and even an epistemological fallacy. Based on Joseph Weizembaum's ideas on the topic, we thus advocate for a strong focus on *human responsibility* in the use of AI.

Keywords: ethics, artificial intelligence, transfer of responsibility, algorithms

*Associate Professor, ESCP Business School

Which tasks shouldn't we delegate to Artificial Intelligence?

Since the early years of artificial intelligence (AI), several examples have showed the risks of an inappropriate use of it. These examples gave raise to important debates since at least the 1970s, during the first wave of AI (usually called rule-based AI) about which tasks we should delegate to AI and which we should not, even if it is technologically possible. Already important at that time, these issues have returned even more strongly with the new wave of AI, which is based on neural networks and machine learning and has led to amazing results, the latest example being ChatGPT and other products of Generative AI. Although digital technologies have developed significantly over the last 50 years, its social, political and ethical issues did (surprisingly?) remain the same. In our case, the questions that still arise are: can we imagine an AI that respects our moral values? What limits should be imposed to the use of artificial intelligence and how to implement these limits?

In this paper we'll first describe some facts that made us worry about certain uses of artificial intelligence, facts that gave rise to an increasing amount of research into the ethical and political issues surrounding AI. We will then look at the two main schools of thought on this subject and conclude by taking a clear position in this debate and making recommendations for the future.

Facts that made us worry

Fact 1: Joseph Weizenbaum introduces ELIZA

In the late 1960s, Joseph Weizenbaum, one of the pioneers of computer science and artificial intelligence, developed the first conversational robot at MIT. This artificial intelligence program (which he called ELIZA) simulated a session with a psychiatrist. You're on the couch, the psychiatrist lets you talk from a very open-ended initial question, and then just bounces off what you say. The programme was very well done and the illusion was perfect: ELIZA was "talking" to you and bounced off what you typed on the keyboard. Weizenbaum introduced this program to some psychiatrists and psychoanalysts in order to show that a machine could not really imitate a human being. He was surprised when he saw many of them delighted to see ELIZA working as if it were a real psychiatrist, and even promote its use to develop psychiatry and psychoanalysis on a large scale and at low cost. Weizenbaum reacted by calling on psychiatrists and psychoanalysts: How can you imagine for a moment to delegate something as intimate as a session with one of you to a machine? (Weizenbaum, 1976)

Fact 2: Stanislav Petrov saves the world in September 1983

On the night of 25-26 September 1983, Stanislav Petrov was the duty officer at the command centre of the Soviet nuclear early-warning system. At 00.15 a.m., Moscow time, the computerised missile warning system alerted to one and then four Minuteman 3 intercontinental ballistic nuclear missile launches from the United States. These launches had been detected by a Russian early warning satellite. Petrov had only a few moments to analyse the situation. Because the number of missiles detected was so low, he disobeyed procedure and told his superiors that he thought it was a false alarm (normally, a nuclear attack should involve dozens or even hundreds of nuclear missiles). Fortunately, his advice was followed and thus avoided a Soviet retaliation that could have been the beginning of a

nuclear war between the Communist countries and the Free World. It was subsequently determined that the false alarm had been created by a rare alignment of sunlight on high-altitude clouds above North Dakota and the orbits of the soviet satellites, which led the AI-based software in the satellites to misinterpret the reflection of sunlight off clouds as the release of energy when nuclear missiles are launched. ("1983 Soviet nuclear false alarm incident", Wikipedia, 2023)

Fact 3: The DoD proposes an AI-based system to make nuclear missiles autonomous

By the end of 1983 (a year of all dangers), the DARPA, the US military research agency, proposed a project in which weapons systems would be equipped with AI-based functions equivalent to those of human beings in the field of autonomous reasoning. Of particular relevance was the plan for a "co-pilot" to assist the flight crew of fighter jets or bombers, who would not only relieve the crew of routine operations, but could also supplant them in missions requiring the ability to accept high-level instructions or for decisions that could not be made on the basis of consensus.

Department of Defence (DoD) experts confirmed this approach - of excluding the human factor from decision-making - in relation to a space-based laser designed to cripple Soviet long-range missiles in their launch phase. Indeed, when reporting this project to the US Congress, the DoD experts acknowledged that this operation would require such a quick action that it would have to be computer programmed, ruling out any White House intervention.

Of course, this statement sparked a strong debate between the DoD and some representatives and senators. Of particular relevance was the DARPA Director's response to a senator's question about whether a programming error could cause the Soviets to launch a real attack. "The President can make a mistake, while we could have a technology that would make no mistakes!" (Desbois, 1985)

Fact 4: In the USA, artificial intelligence helps to deliver justice

In 2017, the Wisconsin State Supreme Court sentenced Eric Loomis, a repeat offender with a criminal record, to six years in prison, and this conviction was made at least in part on the recommendation of a proprietary (and secret) software program from a private company (Northpointe Inc). Loomis later claimed that his right to a fair trial was denied because neither he nor his lawyers were able to review or challenge the algorithm behind the recommendation.

The report suggesting that Loomis be convicted was produced by an AI-based software program called Compas, which is marketed and sold by Northpointe to the courts. This program is one incarnation of a new trend in artificial intelligence: one that aims to help judges make "better" decisions. The judges of the Wisconsin Supreme Court sentenced this repeat offender, adding that the report produced by the Compas software had provided valuable information to their decision, but qualified this by saying that he would have received the same sentence without the report. (Markou, 2017)

Fact 5: Artificial Intelligence fires people

In 2017 and 2018, Amazon used AI software to recruit, evaluate, promote and even fire hundreds of employees (deliverers), with no human being involved in the process. In 2021, Xsolla, a Russian software company specialised in payment solutions for online gaming, fired 150 of its 500 employees. During the pandemic, the company tracked its employees

who were teleworking and then used Big Data and AI to decide who should be fired. On August 17, 2022, 60 Accenture employees who were working for Facebook in Austin, Texas, were "randomly" fired by an artificial intelligence program (Durand, 2022).

Not surprisingly, all these examples (and many others) have sparked important political and ethical debates, but have also opened up a whole new field of research. The field covered by the links between ethics and artificial intelligence is already quite extensive (Haenlein et al., 2022). On the one hand, there is the ethics that we should "inject" into artificial intelligence programmes, and on the other hand, the ethics of the use of artificial intelligence, i.e. making decisions about the tasks that can be delegated to it and those that shouldn't.

Ethics *within* artificial intelligence

The examples described above show that artificial intelligence systems are vulnerable to errors introduced by their human creators. Moreover, the data used to train these artificial intelligence systems may themselves be biased. For example, the facial recognition algorithms created by Microsoft, IBM and Face++ were all biased when it came to detecting gender: these AI systems were able to detect the gender of white men more accurately than colour-skinned men. Another example of racial bias is that of an Asian traveller who was rejected by the software used at the checkpoint because the program said the traveller's eyes were "closed". Amazon stopped its use of AI for hiring and recruiting because the algorithm favoured male applicants over women. This is because Amazon's system was trained with data collected over a period of 10 years and mainly from male candidates.

The issue could therefore be relatively simple: the biases of artificial intelligence algorithms simply translate those of humans. These biases need to be addressed to solve the problems. To stay with the examples above, it would be enough to train the facial recognition algorithms created by Microsoft, IBM and Face++ with more coloured people and Asians. And the best way to correct the bias in the AI system for hiring at Amazon would be to train the algorithm with an equal amount of data from women and men.

But biased data is not the only problem. In addition to that, "data does not capture everything about most real problems. Data is a proxy of reality, which usually is much more complex. In particular, data cannot capture the current context (e.g., what is happening right now) nor what will happen in the future (e.g., all possible traffic accidents in the future)" (Baeza-Yates and Villoslada, 2022). Finally, thanks to machine learning and neural networks, machines can train themselves on the data they receive, and sometimes the results are... random. Microsoft experienced this painfully when it launched its Twitter chatbot Tay in 2016. In less than 24 hours the chatbot went from "Humans are super cool" to "Hitler was right. I hate Jews".

Machines, because they are machines, can never behave ethically because they cannot imagine what a "good life" would be and what it would take to live it. They will never be able to behave morally *per se* because they cannot distinguish between good and evil. The limitations of teaching and algorithm to understand right and wrong should warn against overconfidence in our ability to train them to "behave" ethically.

Ethical *use of* artificial intelligence

In his seminal and influential book, Computer Power and Human Reason, Joseph Weizenbaum poses an essential question: Are there ideas that will never be understood by a machine because they are related to goals that are inappropriate for machines? This

question is essential because it goes to the core of the existence (or not) of a fundamental difference between human beings and machines. For most of us, the existence of a difference is still obvious, but with the development of the new digital culture and the progress of artificial intelligence, it is not clear that this difference will continue to be obvious. Indeed, if the new generations are trained in the idea that human beings are information processors, in the explanation of the mind in terms of neural networks and algorithms, where will they see a fundamental difference between human beings and machines? In the late 90s, when IBM's Deep Blue computer won over Garry Kasparov, the then world chess champion, the question came up again. Asked what could now differentiate a human being from a machine, philosopher John Searle said: "You know what the difference is between Kasparov and Deep Blue? The difference is that Deep Blue doesn't even know that it beat Kasparov." And, we might add, it doesn't occur to him to call up other computers to go for a drink and celebrate its victory.

For professor Searle, the main difference between a human being and AI is consciousness (Searle, 2014). No machine can be conscious, by definition. Weizenbaum argues that the comprehension of humans and machines are of different nature. Human comprehension is based on the fact of having a brain, but also a body and a nervous system, and of being social animals, something a machine will never be (even if social robotics is undergoing significant development nowadays, something that Weizenbaum imagined nearly 50 years ago). The basis on which humans make decisions is totally different from that of AI. The key point is not whether computers will be able to make decisions on justice, or high-level political and military decisions, because they probably will be able to. The point is that computers should not be entrusted to perform these tasks because they would necessarily be made on a basis that no human being could accept, i.e. only on a calculation basis.

The fundamental ethical issue of AI thus seems to us to be the dilution of responsibilities, the transfer of responsibility from the human being to the machine ("I didn't kill her, it was the autonomous car!", "I didn't press the nuclear button, it was the artificial intelligence!"). Even if in the European Union the GDPR (General Data Protection Regulation) is supposed to prevent decisions about humans being made by a computer, we know how things are in justice administrations and HR departments: people are always overwhelmed, and they will not take the time to discuss the advice gave by artificial intelligence ("Nothing personal, Bob; we just asked the AI and it said that you should be fired. But we made the decision!").

Therefore, we believe that Weizenbaum is right: "If the entirety of human experience and the belief structure it entails cannot be formalised, then there are goals appropriate for humans that cannot be appropriate for machines. And if one were to conclude... that there are indeed such goals, then one could say something about what should or should not be delegated to machines." (Weizenbaum, 1976) These issues cannot be addressed by questions that start with "can we?" The limits we must place on the use of computers can only be stated in terms of "should we?" The facts described and analysed here show that since we currently don't know how to make computers wise, we should not delegate them tasks that require wisdom.

Conclusion

With the new generation of robotics and artificial intelligence, a new race for "smart" weapons has begun. The great danger is to have killer robots (or drones) that can make the decision to kill independently of any human decision. The danger is real, to the point that several renowned experts in artificial intelligence signed an open letter in 2015 on the dangers of autonomous weapons, published by the Future of Life Institute. This open letter

was also signed by Elon Musk (the founder of Tesla and SpaceX), Steve Wozniak (the co-founder of Apple with Steve Jobs) and the astrophysicist Stephen Hawking, while he was alive. (Future of Life Institute, 2015)

In recent years, there have fortunately been several initiatives to regulate developments in AI. The European Union, UNESCO, the OECD, the Association for Computer Machinery (ACM), have published documents in this sense. Last but not least, following the amazing developments of Generative AI (ChatGPT, Dall-E, etc.), a new open letter was published in March 2023 by the Future of Life Institute, warning against "an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control" (Future of Life Institute, 2023). The verb used is always "should" (e.g., "Should we risk loss of control of our civilization?"). The signatories call "on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium."

Let us recall the questions we asked at the beginning of this paper. Regarding the question "can we imagine an AI that respects our moral values?", the answer is "no". Rather than trying to teach algorithms to "behave ethically", the real issue is: *Who is responsible here?* We believe that we should focus on the ethical use of AI, on questions like "which tasks shouldn't we delegate to artificial intelligence?" To begin answering this question, we will follow Joseph Weizenbaum who said that since we don't know how to make artificial intelligence wise, we shouldn't delegate it tasks that require wisdom.

More precisely, "which limits should be imposed to the use of artificial intelligence and how to implement these limits?" Some recommendations have been made in many European countries However, we think that it is necessary to be more concrete and therefore more research is needed to answer the latter question, especially how to control the limits that will eventually be imposed to AI. In this sense, how to create a power coalition that can impose these limits is probably the key question.

References

Baeza-Yates, R. and Villoslada, P. (2022), "Human vs Artificial Intelligence", Paper presented at the *IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*.

Desbois, D. (1985), "Comment l'intelligence artificielle conduirait la guerre", *Le Monde Diplomatique*, September 1985.

Durand, K. (2022), "Recrutement, licenciement... Quand l'intelligence artificielle prend la direction des ressources humaines", *Le Figaro*, 29-08-2022.

Future of Life Institute (2015), "Autonomous weapons: an open letter from AI & Robotics researchers" (https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1&cn-reloaded=1&cn-reloaded=1)

Future of Life Institute (2023), "Pause giant AI experiments: an Open Letter" (https://futureoflife.org/open-letter/pause-giant-ai-experiments/)

Haenlein, M., Huang, MH. and Kaplan, A. (2022), "Guest Editorial: Business Ethics in the Era of Artificial Intelligence", *Journal of Business Ethics*, Vol. 178, Issue 4.

Markou, C. (2017), "Why using AI to sentence criminals is a dangerous idea", *The Conversation*, 16-05-2017.

Searle, J.R (2014), "What your computer can't know", *New York Review of Books*, October 2014.

Wallach, W. and Allen, C. (2009), *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, Oxford.

Weizenbaum, J. (1976), *Computer Power and Human Reason: from Judgement to Calculation*, W.H. Freeman and Company, San Francisco and London.