Chapter 4
In AI We Trust: power, illusion and the control of predictive algorithms

Helga Nowotny

1. Introduction: what is AI? The societal context of digital technologies

Intense and accelerating involvement with digital technologies is having a profound impact on the ways in which we work and live, transforming our societies and economies. It challenges us to invent novel ways to use digital technologies to advance the common good instead of mainly increasing the concentration of economic power in the hands of the few. This entails unlocking the great potential of digital technologies to meet the impending risks of climate change, the next pandemics and other emergencies over the horizon. In the workplace, digital technologies continue the processes of automation that began a long time ago, complementing and replacing an ever larger range of tasks, skills and professional activities. Hence, the challenge before us is how to guarantee the development and deployment of a technology – epitomised by artificial intelligence – in a way that makes it sufficiently responsive to the human needs and rights that are being redefined in the process. This includes allocating accountability and responsibility within a complex association between the users of digital technologies and the designers, producers, owners and regulators of the latter.

As have other technologies previously, digital technologies raise questions about whether machines will eventually control, dominate or even fuse with humans, raising fears about dehumanisation, surveillance and totalitarian control. On one side, techno-utopians hail the disruption as bringing the solution to all problems, a 'liberating' force that would even lead to 'greater world harmony' (Negroponte 1998). On the other side are pessimists with their dystopian visions, heralding the end of humanity. Before entering this discussion, we should remind ourselves of the more nuanced relationship between technologies and the humans that design and use them. As David Nye, the technological historian, observed some time ago 'artefacts emerge as the expression of social forces, personal needs, technical limits, markets and political considerations' (Nye 2006). The digital gadgets, infrastructures, networks and machines that serve us now, and that we serve through our behaviour, are not predetermined once and for all. Rather, it is up to us to appropriate, modify and shape them through the choices we make, albeit within societal and technological constraints.

These are some glimpses of the societal context of digital technologies that must be kept in mind when answering the question: what is AI? Today, we have arrived at a crucial point in a long, unprecedented, evolutionary journey marked by the entanglement of multiple interactions between humans and digital machines. Its beginning dates back to the 1940s, but it was only around the first decade of the twenty-first century that a

convergence of three different strands unleashed the power of artificial intelligence that we are witnessing today: the enormous increase of computational power that enables sensors and computer chips of miniature size to be installed in almost every device and everywhere; the development of ever more sophisticated algorithms; and, last but not least, access to and the increasing availability of an enormous amount of data coming from many fields of application.

Therefore, the definition of AI varies according to where we find ourselves in this trajectory. In the beginning, a naïve definition predominated, built on the mathematical-formal approach that started with Alan Turing's insights into the possibilities of developing a mathematical code to run a machine (Turing 1936). This opened the gate to a world in which, as per the definition of the Turing test, AI would simulate human intelligence in machines programmed to think like humans and mimic their actions, whereby 'thinking' was largely equated with formal reasoning. Mathematical code needed hardware to operate electronically in a computing machine and, spurred by the war effort in the 1940s, the technological progress of computers and their performance quickly advanced. The term 'artificial intelligence' was coined at the Darthmouth Conference in 1956 and, even if it is not the most fortunate term, it is still with us. It invites different meanings and interpretations of what 'intelligence' is, especially when juxtaposing human intelligence with the very different 'intelligence' of a machine.

This ambitious but naive definition became replaced by a more realistic one as the formal logical approach failed to yield many of the hoped for practical applications. A decline of funding set in, followed by a period remembered as the 'AI winter'. The decisive turn came at the beginning of the twenty-first century when neural networks began to be deployed, capable of discovering patterns and statistical correlations in data with astonishing efficiency and accuracy. This led to the rise of procedures termed 'machine learning' and 'deep learning'. Algorithms are trained for pattern recognition with the help of large amounts of data capable of training themselves in an approach referred to as 'unsupervised learning'. Hence, AI is defined as any agent or system that perceives its environment and takes actions that maximise its chance of achieving its goals. This implies that a machine must be able reliably to recognise patterns in the environment in which it is expected to act, as with automated vehicles. The goal needs to be defined in precise ways, for instance when following and exploring the rules of games like chess or Go which have resulted in spectacular demonstrations of AI defeating the world's best players.

A third speculative definition of AI has pervaded the public discourse as part of the attempt to realise artificial general intelligence. This entails the still hypothetical ability of an intelligent agent to understand and learn any task that a human being can do. In this definition, an AI would be able to achieve a kind of superintelligence through recursive self-improvement, leading to a point called 'the singularity' by inventor Ray Kurzweil in which the AI will overtake human intelligence. Not surprisingly, this would fundamentally change what it means to be human. While some techno-utopians celebrate this as the ultimate feat of overcoming our humanity by reaching transhumanism, for others this poses an existential risk and the end of humanity as we know it.

2. The power of predictive algorithms: where it comes from and how it affects us

Wanting to know what the future holds has been an ardent wish in all civilisations we know, resulting in divination practices to be found everywhere. Ancient Chinese oracle bones show cracks on the shoulder blades of sheep or turtles that had been held over fire by divinatory experts in order to 'read' the future. Today, we resort to foresight reports and analysis of future trends. While the tools have changed, we are as keen as our ancestors to learn what to expect in the decades ahead and rely increasingly on predictive algorithms. They allow us to build simulation models that answer the question 'what if?' and to expand our imagination while engaging in future-making.

In this process, financial markets for example underwent an intense phase of computerisation and the rapid evolution of automated computer algorithms. Developed by humans, the actual decisions to buy and sell are made by the algorithms and executed through a comprehensive digital infrastructure that links individual trading firms to the various exchanges on which they trade (MacKenzie 2021). The accuracy of weather predictions has also increased dramatically, enabling the worldwide transportation and mobility networks we rely upon today. The trajectory of a hurricane can be followed in real time and its landfall predicted, providing more time for evacuations. Another rapidly expanding field with more dire consequences are automated weapons systems, including drones deployed for military purposes but which also have commercial applications.

Predictive algorithms work not only for governments, the military and business, but for all of us. We rely on them as individuals when wanting to know our future state of health and the risks carried through our genes or lifestyle. We use predictive algorithms for everyday decisions that facilitate our work and personal choices. We collude with the large digital corporations when we feed them our personal data, transforming ourselves into 'the product' that is then sold by them to advertisers who target us in return for the convenience of receiving their services. As Shoshana Zuboff has shown in impressive detail, we have become part of the surveillance capitalism that thrives on the widespread use of predictive algorithms (Zuboff 2019).

Decision-making based on predictive algorithms rapidly pervades not only the business world but public institutions like the police, judiciary, education and the health system. The line to tread between a desirable increase in efficiency and threats to privacy is thin and needs to be continuously re-negotiated within a firm regulatory framework. An example of the trade-offs often involved comes from Arbeitsmarktservice (AMS; the Austrian Public Employment Service) which decided to install an algorithm dividing employment seekers into three groups according to the 'objective' criteria of their profile's prediction of their chances of finding employment. This was followed by a public outcry as the establishment of the category of the least employable was deemed to be socially unjust. Despite assurances from the AMS that this group would also have their needs looked after, the algorithm had to be withdrawn. Although the criteria of 'transparency' had been fully met, the demand for social justice prevails, at least for now.

There are more risks that come with predictive algorithms. As I describe in my book *In AI We Trust. Power, Illusion and Control of Predictive Algorithms*, by transferring ever more agency to an algorithm we tend to believe what it predicts and forget that the algorithm is based on an extrapolation of data from the past. All predictions are based on probabilities as the future remains inherently uncertain. When human behaviour follows a belief in the predictions, self-fulfilling prophecies result and this may herald a return to a deterministic worldview. At the heart of our trust in AI lies a paradox: we leverage control over the future and uncertainty while, at the same time, the performativity of AI and the power it has to make us act in the ways it predicts reduce our agency over that same future (Nowotny 2021).

3. Keeping humans in the loop: towards a digital humanism

The digital devices that surround us and with which we continuously interact provide us with feedback and answers to our questions but also nudge us in pre-set directions. A myriad of sensors and digital tools are automating the ways in which business is conducted, cutting costs and increasing efficiency. While the automation of work is not new, nobody knows how fast new jobs will be created to replace the ones that are vanishing. There are benefits to be gained, but they are unequally distributed. Other serious downsides loom as AI reinforces existing power structures and their concentration into monopolies and oligopolies. Biases in society are transferred to the data and to the algorithms on which decisions rely. These may cause harm and support discriminatory practices, with injustice becoming ingrained, in the absence of institutionalised mechanisms to appeal to human judgment. We are currently witnessing the rampant abuse and misuse of AI in spreading 'fake news' and the threat to liberal democracies.

The unanimous response has been the call to develop an ethical or beneficial AI, expected to fulfil a series of criteria like transparency, explainability, responsibility, fairness and more. However, the problems in implementing these legitimate demands are considerable. First, no consensus exists on the ethical principles themselves, as shown in a study of the ethical guidelines issued by governments and leading corporations worldwide (Jobin et al. 2019). Second, attempts to insert ethical procedures into the design of algorithms, like making self-driving cars 'safe', encounter technical problems as no single interface exists to make an algorithm 'see' like a human driver. Third, the focus on the behaviour of entire ethical systems by auditing the outcome, for example whether human rights are being respected, requires implausible specificity (Danks 2022). Ethics, in line with such a conclusion, is necessary but not sufficient. It is not a checklist and its implementation remains open.

What could work instead? Undoubtedly, more regulation is necessary although it is difficult to achieve at international level with Europe poised between the US and China. A movement devoted to digital humanism is attempting to integrate a human-centred approach in the design, production and deployment of AI throughout their systemic interlinkages (Werthner et al. 2022). It seeks to identify specific points of intervention and to be attentive to actual practice in various domains, as well as becoming part of

the education system. The values on which digital humanism is based will be crucial for shaping the future of work and of liberal democratic societies.

The co-evolutionary journey between humans and digital machines has only begun. AI has considerably expanded human capabilities and opened new spaces of knowledge. It is a powerful technology created by humans and therefore it is social. We can use it to build the society we wish to live in by designing and using AI for the common good.

References

Danks D. (2022) AI ethics as translational ethics. https://www.youtube.com/watch?v=UBo5wVs2qgs

Jobin A., Ienca M. and Vayena E. (2019) The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

MacKenzie D. (2021) Trading at the speed of light: How ultrafast algorithms are transforming financial markets, Princeton University Press.

Negroponte N. (1998) One-room rural schools, Wired, 6 (9). https://web.media.mit.edu/~nicholas/Wired/WIRED6-09.html

Nowotny H. (2021) In AI we trust: Power, illusion and control of predictive algorithms, Polity Press.

Nye D.E. (2006) Technology matters: Questions to live with, MIT Press.

Turing A. (1936) On computable numbers, with an application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, S2-42 (1), 230-265. https://doi.org/10.1112/plms/s2-42.1.230

TU Wien (2019) Vienna manifesto on digital humanism. https://dighum.ec.tuwien.ac.at/dighum-manifesto/

Werthner H., Prem E., Lee E.A. and Ghezzi C. (eds.) (2022) Perspectives on digital humanism, Springer.

Zuboff S. (2019) The age of surveillance capitalism: The fight for a human future at the new frontier of power, PublicAffairs.

All links were checked on 23.01.2024.

Cite this chapter: Nowotny H. (2024) In AI We Trust: power, illusion and the control of predictive algorithms, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.