Article

Wilfried
Martens Centre
for European Studies

European View
2021, Vol. 20(2) 182–190
© The Author(s) 2021
DOI: 10.1177/17816858211059249
journals.sagepub.com/home/euv
SAGE

The ethical dimension of human–artificial intelligence collaboration

Michał Boni

Abstract

The development of artificial intelligence (AI) has accelerated the digital revolution and has had an enormous impact on all aspects of life. Work patterns are starting to change, and cooperation between humans and machines, currently humans and various forms of AI, is becoming crucial. There are advantages and some threats related to these new forms of human–AI collaboration. It is necessary to base this collaboration on ethical principles, ensuring the autonomy of humans over technology. This will create trust, which is indispensable for the fruitful use of AI. This requires an adequate regulatory framework: one that is future proof, anticipates how AI will develop, takes a risk-based approach and implements *ex ante* assessment as a tool to avoid unintended consequences. Furthermore, we need human oversight of the development of AI, supported by inter-institutional partnerships. But first we need to create the conditions for the development of AI digital literacy.

Keywords

Artificial intelligence, Human–machine collaboration, Ethical AI, AI digital literacy, *Ex ante* assessment, Future-proof regulation

Introduction

Human–machine collaboration is one of the key challenges of the digital age. The success of this cooperation will bring many advantages for both humans and technology. Modern machines that use artificial intelligence (AI) have the ability to be much better adjusted to humans' needs, and consequently to work much more effectively. As a result,

Corresponding author:
M. Boni, Wilfried Martens Centre for European Studies, 1st floor, 20 Rue du Commerce, Brussels, B-1000, Belgium.
Email: miboni54@gmail.com

models of societal collaboration can be adjusted, leading to greater efficiency and productivity in many areas. Viewed from this perspective, the European public discourse on the role of AI is inspiring.

The many potential uses of AI systems mean that we can find new areas of collaboration, in which ‘trustworthy AI’ will allow humans to maintain control over the technology and build a healthy coexistence. According to the European Commission’s High-Level Expert Group on AI, ‘trustworthy AI’ refers to AI that is lawful (based on regulations), ethical (follows ethical principles) and robust (does no harm) (European Commission 2019). We can build either a coexistence between humans and AI in which humans do not have control and do not fully understand the principles of operation, or a coexistence based on conditions and principles that create trust.

One analysis of the role of AI in society highlights citizens’ views on the impact of AI (Funk et al. 2020). On average 53% of the population view AI as a good thing, while those who view it as a negative development constitute 33%. However, there are many differences in opinions between continents. These views also relate to opinions about the impact of robotics on job automation: according to Funk et al. (2020), ‘(a) median of 48% say such automation has mostly been a good thing, while 42% say it has been a bad thing’.

These opinions could be a reference point for public discussions about AI development and the potential opportunities for a new relationship between humans and machines. The aim of these discussions would be to minimise feelings of being at a disadvantage that those in a variety of roles may harbour. AI systems can and will affect everybody, everywhere, and this should be taken into account when working on the regulatory aspects of AI in the EU.

This article argues that human–machine/AI collaboration must be based on ethical principles. The expected framework for this is through management by law and, at the same time, management by oversight. The latter gives responsibility for the oversight of all AI development processes to partners (in a collaborative way), including science, business, public institutions and civic organisations, with the aim of encouraging innovation within rules that uphold fundamental and consumer rights. The solutions discussed and implemented in the EU can build future competitive advantages for the Union. In the subsequent sections, this article will review the key drivers of the discourse on the AI ecosystem, and highlight the problem of future-proofing legislation with guarantees of equilibrium between human autonomy and technological innovation. In addition, it will define AI digital literacy, describe the significance of *ex ante* impact assessments[1] in ensuring the ethical behaviour of AI functionalities and offer useful recommendations for policymaking in the area of AI development.

The AI ecosystem discourse

It is important to note that work on AI development has been undertaken in a holistic manner, taking into account a wide variety of factors, including science and business measures (i.e. creating an ecosystem of excellence) and trust building as an important social challenge. To establish an ecosystem of trust is a crucial challenge (European Commission 2020a; 2021a; 2021b; 2021c). This ecosystem should be focused on human values, rights, societal needs, aspirations and ensuring that AI works effectively for people. This theme is clearly presented in the annex to the Communication *Fostering a European Approach to Artificial Intelligence* (European Commission 2021b).

How then, can we make AI systems trustworthy and ensure that AI works with people and for people without having negative impacts and causing concern? Certain features are essential for fruitful collaboration between AI and humans, especially regarding the ethical development of AI.

First, AI is fundamental for the development of certain sectors and for their improved efficiency. AI systems have started to change, *inter alia*, the manufacturing and service industries, financial services, the workplace, social care, education, healthcare, security systems, management strategies, climate and environmental policies, administrative services, migration and asylum policies, agricultural policies and mobility solutions. In all these areas, a new model of relationship between humans and robots/AI is starting to be established. And the essential factor in this new relationship is *trust*. This is especially important with regard to the changing ways of working (e.g. new collaborative models, shift arrangements and job sharing, including between humans and robots/AI) with regard to the emotional aspects of work.

Second, it is necessary to understand what kind of regulatory framework will be most suitable for developing AI systems. In many areas, the traditional view is focused on a static regulatory framework (Evas 2020), which is sufficient to solve some problems of harm (e.g. mental, physical or material) potentially caused by humans (e.g. in car accidents, through violence against women and children etc.). Currently, a crucial debate is taking place on the responsibility of social media platform providers for the dissemination of illegal content and disinformation, and the phenomenon of hate speech which have arisen on social networks. In this case it is clear that we can use the universal norms incorporated into regulatory frameworks to indicate challenges and help people tackle problems. This also offers consumers redress mechanisms to allow space for the implementation of rights and to deliver remedies.

In many scientific and political debates in which the significance of AI as a game changer is strongly emphasised, the problem of the possible autonomy of AI is raised. This requires a different kind of response from the legislative side. One option, adaptation, would mean the adjustment of existing legal principles in order to apply them to the functioning of AI systems, as in the case of the EU's Machinery Directive (European Parliament and Council 2006; Tuominen and Festor 2021). The second option, anticipation, would mean that existing legal systems and new proposals need to be able to provide dynamic legal mechanisms to safeguard against new risks (Evas 2020). Future-proof solutions are necessary, and the only way to avoid unintended consequences is through *ex ante* impact assessments. There are currently some obligations regarding *ex post* impact assessments, that is, those that are carried out after the introduction of new solutions (including technological ones)—for instance, in order to understand how new devices and functionalities are working, both in terms of being in accord with the legal framework and their influence on people. Including in the regulations the obligation to conduct *ex ante* impact assessments would be an innovation, but it would be a more effective way of eliminating certain threats.

Third, it is important to find a dynamic model for the regulatory framework for AI due to the dynamism of AI's developing functionalities. The concept of the 'risk pyramid' presented by the European Commission (2021a) is the best solution. For concretely high-risk AI systems, it would establish the requirement to fulfil all obligations of the conformity assessment before market entry. This would mean that all decisions assessing the level of possible threat to humans would use a 'risk-based approach', supported by evidence, not by intuition.

One of the key components from the ethical perspective is the model of the *ex ante* impact assessment. It should be based on full respect for human autonomy: dignity and agency are unique attributes of human beings, and form the foundation of human rights. In this respect, it must be noted that 'to manage and decide about humans in the way we manage and decide about objects or data, even if this is technically conceivable' would be inappropriate (European Group on Ethics in Science and New Technologies 2018, 9). A key factor in the development of AI systems is, of course, data-related: the quantity that is available as well as the quality of it. The normative assumption ought to be clear: humans should not be reduced to the data dimension.

Fourth, it is important to defend human autonomy by establishing that the actions of AI systems can only be based on principles which guarantee respect for fundamental and consumer rights, and by having a transparent discussion about the anticipation of AI systems development and innovation in maintaining values. We need to define all examples of potential breaches of ethical principles. There are many possible threats to human autonomy that could have adverse impacts on a plurality of people, in all kinds of applications (European Commission 2021a; 2021d)—the use of AI that embodies these threats is considered 'high risk' and some types of AI are banned in the European Commission's AI regulation proposal. These include, *inter alia*, the use of autonomous weapons, the presence of AI systems in the management of critical infrastructure, the use of AI to automate certain decision-making processes (e.g. the allocation of social benefits and loans), mass surveillance technologies and the use of AI in law enforcement (Dumbrava 2021).

A lack of ethical principles could undermine the relationship between humans and AI. All the reviewed cases and AI uses clearly show an increased risk to citizens' fundamental rights and the potential for the violation of EU values; threats to human dignity and the violation of personal autonomy; concerns over a reduction in privacy protection despite the existence of the EU's General Data Protection Regulation; the visible growth of discriminatory outcomes led by algorithms; and the operation of 'black box' models which do not use explainable mechanisms. These examples show the unbalanced relationship between humans and AI.

Fortunately, the European Commission's proposal on AI regulation, which would establish the risk pyramid concept and conformity assessment procedures, aims to minimise some risks and limit some dangerous solutions. The minimal transparency obligations for non-high-risk AI systems presented in the proposal is very valuable. The proposal offers a proper response to the need for an equilibrium between the threats and advantages of AI, based on transparent rules and common principles, and translated into technical and business models for the functioning of AI.

AI digital literacy and awareness

The above-mentioned 'equilibrium' also means creating sustainability for both sides: human and machine/AI. The human side relates to reality and potentiality. Humans have many fears about the unknown world of technology and AI. What is unknown is uncertain, and this uncertainty leads to insecurity. How can we change these negative feelings?

The answer became clear during consultations in 2020 (European Commission 2020c): 90% of respondents indicated that improving skills would be the most important action to prepare people to use AI in a better way (the adoption of a training programme was crucial for 86% of respondents). Stakeholders, especially business actors, also raised the issue of AI digital literacy (Digital Europe 2020), which should be integrated in a comprehensive way into all educational formats (building the potential for adaptability).

In addition, when raising awareness of AI, it is important to know as much as possible about the impact of AI on human psychology and behaviour. In some professions there is the possibility of challenging problems, such as emotional attachments to robots, the loss of the ability to think for yourself and to be introspective, the danger of deception and manipulation, the risk of becoming psychologically dependent on robots, and the unpredictability of forming human–robot/AI relationships, which could lead to violent behaviour under the informal, 'intimate' pressure of the AI device, especially in humanoid form. Yet, there are also opportunities: interactions between humans and AI can change human-to-human interaction models and improve the collaborative competences of humans in relation to other humans through the supportive oversight of AI. Knowledge of these problems and advantages, coming from understanding the new opportunities of AI systems development, should form part of AI digital literacy.

Ex ante impact assessments

In establishing trustworthy human–machine collaboration, we have to emphasise the importance of 'soft' instruments. These will be crucial not only for the better adaptability of humans to the increasing presence of machines, but also for the machine/AI side as a way to adjust to users' needs and expectations. This means that the architects, developers and deployers of AI solutions need to integrate new technologies into their work to guarantee an acceptable standard of operation that is in line with ethical principles.

A significant component of avoiding breaches of fundamental rights and discriminatory solutions is linked to the quality of the datasets which are generated using the power of AI. Using faulty or badly compiled data for algorithmic work can create a significant risk of discrimination, whether intended or unintended. The data training volumes held by developers can include unintended social bias, which is then reproduced and automatically reinforced by AI systems. On the one hand, the response to this should focus on applying proper standards to the data selection used for AI training. On the other, there is a need to establish European guidelines for data usage, which is likely to be the subject of the Data Governance Act (European Commission 2020b) and the Data Act expected in late 2021. How users can be given some tools to control their own data is an additional factor for consideration.

Checking the quality of the datasets used for AI training is fundamental to ensuring its appropriateness: the reference data of all high-risk AI systems must be checked for conformity with fundamental rights, privacy protection and ethical values. The obligation to do this should be placed on both the providers (to ensure compliance with AI requirements, registration, quality and risk management, post-market monitoring and reporting to competent authorities) and the users (human oversight, documentation, completion of a Data Protection Impact Assessment) of high-risk AI systems (European Commission 2021d, 53–4).

During consultations on the European Commission's 2020 AI White Paper, it was agreed to focus on self-assessed *ex ante* analysis and conformity assessments carried out by the developer. To ensure the trustworthiness of AI, various impact assessment models were discussed. One of the key proposals is the introduction of a Human Rights Impact Assessment, which should be carried out and documented (Panoptykon Foundation 2020, 4) for all AI systems as a mandatory solution, as proposed by many civil society and consumer organisations. This assessment regime should introduce a mandatory disclosure scheme, describe the role and tasks for external reviewers (an external report should be required), and increase real engagement from those affected (communities, individuals and civil society groups).

This model of compliance would be very resource-heavy, both for companies and public authorities. Additional proposals were also brought to the table. In cooperation with the European Parliament, several scientists (Van Wyngsberghe, 2020) presented on the need for the implementation of an ethical Technology Assessment and the concept of Data Hygiene Certification.

The requirement for *ex ante* procedures in the European Commission's proposal are paramount to the whole process of the conformity assessment. The more companies and people become involved in *ex ante* measures, the more trust will become the fundament of human–AI collaboration.

Conclusion

It should be clear that to ensure future human–AI collaboration and to make solutions much more visible and oriented towards building trust, we have to find the best instruments for AI management. Some of these should come from the formulation of management by law. However, some should be created as a new tool, that is, management by oversight, constructed and used practically and institutionally by all partners (Boni 2021). In addition is also important to check and support the credibility of the institutions responsible for the evaluation of conformity assessments, registration systems and *ex post* market surveillance, especially at the national levels—wherein will lie the enforcement and supervision of the new rules.

However, the overall objective is to achieve transparent conditions for the human oversight of AI mechanisms and functionalities as the key to future fruitful collaboration between humans and machines, and also human–machine/AI teaming. Such conditions need to be adaptable to new challenges and kept as a constant foundation for the benefit of the humans who will play many roles vis-à-vis AI systems opportunities. These humans will include doctors, who will work in cooperation with AI on personalised diagnosis and therapies; officers in public institutions developing automated decision-making processes for public services (including in terms of the scoring of people, with the risk of discriminatory actions); and city managers, who will use AI systems to make security solutions much more effective in many areas without breaching residents' rights.

In this sense, some policies are needed at the EU level. They should ensure that fragmented solutions (separate national models) are avoided; create a holistic view of AI development (supplemented by data governance); and build future-proof regulations to make European AI growth a reference point (in terms of norms and standards) in transatlantic cooperation. To achieve the best oversight, EU policies should bring together the efforts of all stakeholders (from business, science, civil society and European institutions). At the same time action is needed to build the educational capacity of the workforce and improve AI digital literacy in all societies, and financial investment (both public and private) is needed to develop AI functionalities. Moreover, promotion of AI will be indispensable during the process of implementing the new regulations, especially to convince small and medium-sized enterprises to use all the possibilities AI offers. Finally, political will is going to be needed to bring all debates to an effective conclusion and begin implementation of the new rules. Implementation should be monitored since it will open up opportunities for further current uses and reveal future challenges.

Note

[1] In the EU context, *ex ante* impact assessment refers to 'an attempt to provide, in advance of legislating, a coherent analysis of the reasoning that lies behind, and the foreseeable effects of, any proposed measure or policy initiative' (Dunne and Eisele 2020).

References

Bird, E., Fox-Shelley, J., Jenner, N., Leveloy, R., Weikkamp, E., & Winfield, A. (2020b). *The ethics of artificial intelligence: Issues and initiatives.* European Parliamentary Research Service, PE 634.452. March. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf. Accessed 6 September 2021.

Boni, M. (2021). *Artificial intelligence and governance: Going beyond ethics*. Wilfried Martens Centre for European Studies. Brussels. https://www.martenscentre.eu/publication/artificial-intelligence-and-governance-going-beyond-ethics/. Accessed 6 September 2021.

Digital Europe. (2020). *DigitalEurope comments on the European Commission's AI White Paper*. 12 June. https://www.digitaleurope.org/wp/wp-content/uploads/2020/06/DIGITALEUROPEs-response-to-the-AI-White-Paper-consultation.pdf. Accessed 6 September 2021.

Dumbrava, C. (2021). *Artificial intelligence at EU borders: Overview of applications and key issues*. European Parliamentary Research Service, PE 690.706. July. https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/690706/EPRS_IDA(2021)690706_EN.pdf. Accessed 6 September 2021.

Dunne, J., & Eisele, K. (2020). *Ex-ante impact assessment in the EU*. European Parliamentary Research Service, 13 May. https://epthinktank.eu/2020/05/13/ex-ante-impact-assessment-in-the-eu-2/. Accessed 6 October 2021.

European Commission. (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 6 September 2021.

European Commission. (2020a). *Artificial intelligence: A European approach to excellence and trust*. White Paper. COM (2020) 65 final, 19 February. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065&from=EN. Accessed 6 September 2021.

European Commission. (2020b). Proposal for a Regulation on European data governance (Data Governance Act). COM (2020) 767 final, 25 November. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767. Accessed 6 September 2021.

European Commission. (2020c). *Public consultation on the AI White Paper – Final report*. 30 November. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68462. Accessed 6 September 2021.

European Commission. (2021a). *Annexes to the Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* COM (2021) 206 final, 21 April. https://ec.europa.eu/transparency/documents-register/api/files/COM(2021)206_1/de00000000993583?rendition=false. Accessed 6 September 2021.

European Commission. (2021b). *Fostering a European approach to artificial intelligence*. Communication. COM (2021) 205 final, 21 April. https://eur-lex.europa.eu/legal-content/DA/TXT/?uri=COM:2021:205:FIN. Accessed 6 September 2021.

European Commission. (2021c). *Impact assessment accompanying the Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Staff Working Document. SWD (2021) 84 final, 21 April. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=SWD:2021:84:FIN. Accessed 6 September 2021.

European Commission. (2021d). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM (2021) 206 final, 21 April. https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2021)206&lang=en. Accessed 6 September 2021.

European Group on Ethics in Science and New Technologies. (2018). *Statement on artificial intelligence, robotics and ‘autonomous’ systems*. 9 March. https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1. Accessed 6 September 2021.

European Parliament and Council. (2006). Directive 2006/42/EC on machinery, and amending Directive 95/16/EC (recast). OJ L157 (9 June), 24. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32006L0042. Accessed 6 September 2021.

Evas, T. (2020). *European framework on ethical aspects of artificial intelligence, robotics and related technologies: European added value assessment*. European Parliamentary Research Service, PE 654.179. September. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/654179/EPRS_STU(2020)654179_EN.pdf. Accessed 6 September 2021.

Funk, C., Tyson, A., Kennedy, B., & Johnson, C. (2020). Science and scientists held in high esteem across global publics. *Pew Research Center*, 29 September. https://www.pewresearch.org/science/2020/09/29/science-and-scientists-held-in-high-esteem-across-global-publics/. Accessed 6 September 2021.

Panoptykon Foundation. (2020). *Panoptykon Foundation’s submission to the consultation on the ‘White Paper on artificial intelligence – a European approach to excellence and trust’*. Warsaw, 10 June. https://panoptykon.org/sites/default/files/stanowiska/panoptykon_ai_whitepaper_submission_10.06.2010_final.pdf. Accessed 6 September 2021.

Tuominen, M., & Festor, S. (2021). *Revising the Machinery Directive*. European Parliamentary Research Service, PE 694.208. July. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/694208/EPRS_BRI(2021)694208_EN.pdf. Accessed 6 September 2021.

Van Wyngsberghe, A. (2020). *Artificial intelligence: From ethics to policy*. European Parliamentary Research Service, PE 641.507. June. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641507/EPRS_STU(2020)641507_EN.pdf. Accessed 6 September 2021.

Author biography

Michał Boni, PhD, *is a senior research associate at the Wilfried Martens Centre for European Studies and an assistant professor at the Warsaw University for Social Sciences and Humanities. He is a former Member of the European Parliament from Poland and a former Polish Minister of Administration and Digitisation.*