NGI THINK NEXUS

Policy Brief 6

AI Ethics:
A challenge or an opportunity

December 2020

Policy Brief writen by:
Think NEXUS Innovation & Entrepreneurship Expert Group

Think NEXUS, an EC-funded project, aims at reinforcing EU-US collaboration on NGI-related topics in three focus areas: Science and Technology, Innovation and Entrepreneurship and Policy. The aim is to boost strategic research, industrial partnerships and policy compliances in order to gain socio-economic benefits in both the EU and US regions.

In the framework of this project, we are regularly publishing several short articles aiming at comparing the US and the EU approaches in different topics of NGI. The present document is focusing on Artificial Intelligence.

THI

NK

NEX

US

AI Ethics: A challenge or an opportunity

The plethora of policy initiatives, documents, and recommendations related to AI ethics indicates that there is a widely shared belief that AI applications will raise – and are already raising – complex ethical challenges that, if not addressed correctly, may pose great societal risks. After all, AI ethics is not solely about technological change and its impact on individual lives, but also about societal, economic, and cultural transformations and the future of our societies[1].

Even if general or human AI has not been achieved yet, and for some it may never happen in practice, narrow AI is already here and is pervasive, with applications in many sectors, including manufacturing, transportation, agriculture, healthcare, science, education, finance, human resources management, security, entertainment, and marketing. These applications are sometimes visible, as in the case of autonomous cars or cancer diagnostic systems, but, more often than not, they are hidden as parts of larger, more complex systems. Increasingly, different technologies with applications that cover a wide spectrum of human activity will have some AI components embedded in their systems.

This is already happening with drones that use AI, AI conversational agents, AI in the workplace used as part of HR systems, AI that writes articles or AI systems that generate synthetic images and videos of real people, often making them look like saying or doing things they never did. AI is already used in decision making processes in courts, for example the COMPAS system in the US and the HART system in the UK have been used to predict those likely to re-offend. AI is also heavily used for law enforcement purposes, such as predictive policing, facial recognition technologies, behaviour detection, and detection based on biometric data. [2] AI is also permeating more personal and intimate aspects of our life, such as machines that can read our faces and emotions or online dating apps that use AI for matchmaking.

These AI applications raise ethical and legal concerns and create real world risks that directly affect both individual lives and collective societal processes. **Bias in AI**, most often unintentional, may be inserted in all stages of design, testing, and application of an algorithm.

1 Mark Coeckelbergh, AI Ethics (Cambridge, MA, USA: The MIT Press, 2020).

2 European Parliament Committee on Civil Liberty Justice and Human Affairs, Draft Report on Artificial Intelligence in Criminal Law and Its Use by the Police and Judicial Authorities in Criminal Matters, 2020 <https://www.europarl.europa.eu/doceo/document/LIBE-PR-652625_EN.pdf>.

In law enforcement applications for example, research has shown that the use of the COMPAS system has already led to false positives and false negatives[3], reinforcing existing biases and unjust discrimination already present in our societies and thus our data. So bias seems to be ubiquitous in our world and societies, suggesting that AI models will never be totally bias free; although there are still things we can do to minimise bias and take quick corrective measures[4].

The use of facial recognition systems and so called emotional AI enables machines to identify us and read our emotions, collect our biometric data, predict our mental, emotional, and physical status without even being noticed, raising serious **concerns over data protection and privacy**. These concerns are not new, but AI applications magnify the problem, as they collect and process much more personal data, often operating in the background, without users even knowing that AI is used. This may in turn lead to other situations where AI systems are used to manipulate democratic processes and influence voting decisions (e.g. in the case of Cambridge Analytica or social media bots propagating political disinformation).

The nature of some AI algorithms also raises **the problem of the "black box society"**, a term coined by Brooklyn Law School professor Frank Pasquale to express the notion of a networked society based on opaque, nontransparent algorithmic systems[5]. We may know how these algorithms work, but we don't fully understand how they come to a particular decision, and thus we cannot fully explain this decision, albeit the impact the latter may have on an individual's life. Current machine and deep learning applications are too complex in their automated decision making processes. There is the danger that even the people behind the creation and operation of AI systems won't be able to understand every step of the process and explain the algorithm's specific decisions, let alone end users of these systems or policy makers.

These concerns raise the issue of **transparency and explainability of AI systems**; how can we make AI more transparent and explainable, and is this fully possible? The issue of algorithmic transparency or opacity is also closely connected with other issues, such as responsibility (an ethical concern over who should be responsible over automated decisions), liability (a legal concern about who should be held liable for unlawful consequences), and explainability (as a philosophical concern over the very nature of explanation and the human decision making process).

How many decisions, how much of these decisions, and what kind of decisions should be delegated to AI algorithms, especially if we can't fully explain the decisions they make? Should we design AI systems to adhere to ethical requirements and if yes, which requirements and to what extent? Do we agree every time on the interpretation and real world implementation of ethical principles and how do we reconcile contesting ethical views? To what extent will we need or want to balance explainability, privacy, fairness with other considerations like financial benefits, innovation, and competiveness? Who shapes the future of AI?

3 Julia Angwin and others, 'Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.', ProPublica, 2016 <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [accessed 30 November 2020]. For further analysis see Julia Dressel and Hany Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism', ScienceAdvances, 1.4 (2018) <https://doi.org/10.1126/sciadv.aao5580>.

4 Digital Europe, Recommendations on AI Policy: Towards a Sustainable & Innovation Friendly Approach, 2018 <https://www.digitaleurope.org/wp/wp-content/uploads/2018/11/DIGITALEUROPE-Recommendations-on-AI-Policy-November-2018-final.pdf>.

5 Frank Pasquale, The Black Box Society, 1st edn (Harvard University Press, 2015).

The breadth of ethical questions and problems raised in the context of AI has highlighted the need to include ethics in the AI discourse. As such, a wide range of public and private policy initiatives around the world have included ethical aspects into their AI policies; for example, the European Commission set up in 2018 the High-Level Expert Group on Artificial Intelligence, which released a set of ethical guidelines towards developing trustworthy AI[6]. Others talk about explainable AI, while other ethical principles for AI include those of justice, fairness, sustainability, non-maleficence, freedom and autonomy. Although the HLEG AI and other relevant initiatives have formulated many ethics guidelines and some attempt to apply these guidelines through technical guidelines and processes, generally **ethics components in AI policy still tend to remain too vague in their implementation, fall under a soft regulation approach, and are distant from the actual technology environment and practices**.[7]

To better cope with the ethical challenges raised by AI, more proactive and radical policy responses need to be formulated that will enable us to effectively embed ethics guidelines in development and validation processes. To foster a culture of responsible AI innovation that integrates ethics by design and privacy by design approaches, policy makers need to **seek convergence not only on the ethical principles of AI but also on the actual implementation of these principles through processes**. It is also important that this process is as much democratic and inclusive as possible. To this end, a bottom-up approach should replace the current top-down approach in formulating policy recommendations.

This can be achieved through **broad stakeholder engagement** that includes AI researchers, academics, and professionals, people affected by the applications of AI, and citizens; **open public debates and forums**; as well as **early societal involvement and intervention in research and innovation processes**. Although initiatives such as the **European AI Alliance** move towards this direction, it needs to be ensured that these efforts will actually reach the "lowest" levels and actively seek to engage communities like developers and end users of AI systems in their processes, as well as specific population groups that are affected by the use of the AI system in question.

Moreover, we need to **encourage a multicultural and transdisciplinary research and innovation environment** and to **foster ties to exchange knowledge and practices** with the US, but also with non-Western political and cultural systems that may have to offer us new approaches and alternative "AI lessons learned". At the moment there seems to be little collaboration in international level and between the social sciences and the natural sciences when it comes to AI research and development, even within Europe. A more holistic approach requires the involvement of both social and natural sciences in R&D projects and the education of each side's professionals on the other side's main concepts. In other words, philosophers, psychologists, and lawyers need to learn how machine learning works, while software developers, data analysts, and system engineers need to understand the ethical and legal concepts concerning AI.

6 European Commission High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 2019 <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

7 For example, recent criticism of the EU's approach towards AI focuses on the fact that the ethics guidelines for trustworthy AI fail to ensure the effective protection of human rights. See Access Now and European Digital Rights (EDRi), 'Attention EU Regulators: We Need More than AI "Ethics" to Keep Us Safe', EDRi <https://edri.org/our-work/attention-eu-regulators-we-need-more-than-ai-ethics-to-keep-us-safe/> [accessed 30 November 2020].

The matter of education is crucial in general. **More policy focus should be given on designing appropriate education concerning AI**, its uses, and applications in all levels of the society. From schools and universities to professionals in different industries, end users, scientists and policy makers, education should be tailored to the specific context, target group, and objectives of each case. **Algorithmic literacy should be generally enhanced** and people should be familiarized with the technological, ethical, and legal concepts around AI in order to be able to meaningfully participate in public debates and consultations.

Last but not least, to fully materialize the dynamic and possibilities of AI ethics we shouldn't hesitate to **put up into debate broader ethical and philosophical questions** about what is important and valuable for us as individuals and our societies moving to the future and how can we transcript these concepts into AI applications.

For example, should we try to design AI systems following the principle of non-maleficence, as in "you should not cause any harm", or the principle of benevolence, as in "you should also do good things"?[8] In the latter case, we also need to define what good means and for whom. A more positive approach towards ethics will therefore require not only to set ethical constraints on AI systems; but also to actively engage with broader ethical questions and collectively seek answers and ways to shape a vision of the future that reflects these notions.

8 As research shows, currently the principle of non-maleficence is used more often in policy guidelines than the benevolence principle (Anna Jobin, Marcello Ienca, and Effy Vayena, 'The Global Landscape of AI Ethics Guidelines', Nature Machine Intelligence, 1 (2019) <https://doi.org/https://doi.org/10.1038/s42256-019-0088-2>).

NGI THINK NEXUS

thinknexus.ngi.eu @ThinkNEXUS_NGI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825189. This publication reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains