BRIEFING

STOA Options Brief

European Parliament

The ethics of artificial intelligence: Issues and initiatives

In recent years, countries and regions around the world have created an increasing number of guidelines and frameworks to address the ethical and moral implications arising from the development of artificial intelligence (AI). Based on a broader study, 'Exploring ethical concerns and moral questions in the context of artificial intelligence', this options brief identifies crucial gaps that could benefit from being addressed in future legislation, regulation or guidelines (see the accompanying study for further detail). Relevant but under-addressed issues include: the mechanisms of fair benefit-sharing; exploitation of workers; energy demands in the context of environmental and climate change, and the potential for AI-assisted financial crime. This briefing highlights the scope of coverage of these key gaps under the European Commission's 'Ethics guidelines for trustworthy AI' (henceforth 'ethics guidelines'), and presents some options and considerations for addressing them.

Key gaps in the ethics guidelines

Moral and ethical dilemmas, as identified by this study, are addressed to varying extents in the European Commission's Ethics Guidelines. However, there are some notable gaps.

Environment

Specific to AI, machine learning will require more and more data to be processed, which requires huge amounts of energy. The use of AI will also require large amounts of energy for manufacturing and training – for example, it would take many hours to train a large-scale AI model to understand and recognise human language, to a level where it could be used for translation purposes.[1] According to Strubell, Ganesh, and McCallum,[2] the carbon footprint of training, tuning and experimenting with a natural language processing AI is over seven times that of an average human in one year, and roughly 1.5 times the carbon footprint of an average car, including fuel, across its entire lifetime.

Coverage:

The 'Ethics guidelines for trustworthy artificial intelligence (AI)' prepared by the European Commission's High-Level Expert Group on Artificial Intelligence are founded on a principle of prevention of harm and a principle of fairness, which means that systems must be able to operate without harming living beings or the environment. The assessment list includes explicit mention of monitoring energy from data centres and consideration of environmental impact assessment. Particular examples are also given on how to achieve this (e.g. critical assessment of resource use and energy consumption throughout the supply chain). However, it is not explicit whether the guiding principle of prevention of harm extends to resource depletion and energy consumption from unsustainable or polluting sources.

Options:

- Data processing associated with AI consumes high levels of energy. A stronger emphasis on sustainability and environmental responsibility is required in the development of AI systems.
- The gap in understanding of benefits versus costs could be addressed through such means as cost-benefit studies and life-cycle analyses, or costing methods which include environmental externalities.
- Requiring energy use monitoring, identification and publication of company carbon footprints and improved tracking of energy supply chains are possible options for addressing AI developers' energy use.
- There is debate about whether the regulation and incentivisation of technology development should be problem-led or possibility-led. One option is to use laws to 'direct' the course of technology innovation towards urgent environmental priorities. Mission-oriented policy[3] is one approach that could help direct the efforts of technology developers towards societally relevant outcomes.

Inequality

Besides changes to the labour market, two inequality-related ethical issues are inequitable distribution of the benefits (such as profits, influence, data and solutions), and issues around job quality, with highly skilled workers performing repetitive tasks, such as moderating and tagging content or cleaning the huge datasets needed to train the technology.

Sharing the benefits

Some argue that progress in AI will come at the expense of the human workforce. This will mean that particular jobs, and perhaps whole industries, may become redundant, with revenues split across fewer people, increasing social inequalities. Consequently, individuals who hold ownership in AI-driven companies are set to benefit disproportionately. Brundage and Bryson[4] state that 'it is not sufficient to fund basic research and expect it to be widely and equitably diffused in society by private actors'. Issues also arise regarding the accumulation of technological, economic and political power in the hands of the top five players – Google, Facebook, Microsoft, Apple and Amazon – which affords them undue influence in areas of society relevant to opinion-building in democracies: governments, legislators, civil society, political parties, schools and education, journalism, and science and research.

Exploitation of workers and human cost

Better (and safer) AI needs huge training data sets and a whole new outsourced industry has sprung up worldwide to meet this need. This has created several new categories of job – many of which now form part of the 'gig economy', placing workers outside the protection of labour laws. This should be considered when characterising the benefits of AI to society. These include: (i) scanning and identifying offensive content for deletion, (ii) manually tagging objects in images to create training data sets for machine learning systems (for example, to generate training data sets for driverless car AI applications) and (iii) interpreting queries (text or speech) that an AI chatbot cannot understand.

One of the key ethical issues is that – given the price of the end-products – these temporary workers are being inequitably reimbursed for work that is essential to the functioning of the AI technologies. Another issue regards workers required to watch and vet offensive content for media platforms, such as Facebook and YouTube.[5] Such content can include hate speech, violent pornography, cruelty and sometimes the murder of either animals or humans. A news report[6] outlines mental health issues (post-traumatic stress disorder (PTSD)-like trauma symptoms, panic attacks and burnout), alongside poor working conditions and ineffective counselling.

Coverage:

The ethics guidelines include diversity, non-discrimination and fairness as requirements. The guidelines elaborate that equality is a fundamental basis for trustworthy AI and state that AI should be trained on data which is representative of different groups, in order to prevent biased outputs. Preventing harm is covered: in situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. The EU assessment checklist in the ethics guidelines also includes accessibility and stakeholder participation requirements – and 'broader societal impact' – but stops short of asking 'who benefits from the AI?' (benefits could come in the form of profits, data or solutions). Under the transparency section, the EU assessment checklist mentions tracing data, methods or scenarios used to to test the algorithms, but does not mention traceability through the value chain or of workers' inputs. The ethics guidelines mention provision of quality education to ensure the right skills are available to fill the future jobs – however, they are not explicit in mentioning new classes of jobs created, nor mentioning the possible implications of hidden work, harmful jobs or precarious working conditions.

Options:

- One option may be to declare that AI is not a private good, but instead should be available for the benefit of all. [7] This would require a change in cultural norms and policy as well as new strategies to harness the beneficial powers of AI, help navigate the AI-driven economic transition, and retain and strengthen public trust in AI.[8]
- On corporate social responsibility, it may be worth instigating minimum acceptable reporting requirements (as mandated in Directive 2014/95/EU with regard to human rights, environmental and social responsibility) for transnational corporations and large enterprises to show how they are sharing the benefits of their AI technology.
- AI-driven job losses will require new retraining programmes and social and financial support for displaced workers; such issues could require economic policies, such as AI taxation schemes,[9] or targeted, industry-specific retraining. Policies should focus on those most at risk of being left behind.
- Making worker inputs more transparent in the end-product could help to add value to this work and improve the equitable distribution of benefits along the value chain.
- Appropriate support structures and working conditions are needed for precarious workers, and for those working in unhealthy jobs or with psychologically harmful content.

General considerations

- Although the ethics guidelines acknowledge the beneficial use of AI in finance, they do not adequately addresses potential negative impacts on the financial system, either through accidental harm or malicious activity. The potential for AI-assisted financial crime is an important one and currently unaddressed by any international framework.
- Governments and regulators need to develop new forms of technology assessment – allowing them to deepen their understanding of such technologies, while they can still be shaped.
- Robust ethical principles are essential in the future of this rapidly developing technology, but not all countries understand ethics in the same way. A number of countries have committed to creating AI ethics councils.
- Placing the burden of proof on the developer to demonstrate safety, environmental considerations and/or public benefits could help to enforce stringent standards of safety, without prescribing how it is done and stifling innovation.

- Incorporating non-prescriptive systems using a set of principles that need to be adopted, interpreted and issued within a company, gives regulators more scope to question the approach taken and require developers to engage with regulators.
- A single regulatory body, providing prescriptive guidance to national regulators, could help to eliminate incoherent and conflicting sets of standards and guidance.
- Ensuring that value chains are traceable and trackable could be addressed via public procurement rules.

MAIN REFERENCES

[1] A. Winfield, Energy and Exploitation: AIs dirty secrets, June 2019.

[2] E. Strubell, A. Ganesh and A. McCallum, Energy and Policy Considerations for Deep Learning in NLP, 2019.

[3] M. Mazzucato, Mission-Oriented Research & Innovation in the European Union, European Commission, 2018.

[4] M. Brundage and J. Bryson, Smart Policies for Artificial Intelligence, 2016.

[5] S. Roberts, 'Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste', *Media Studies Publications,* 2016.

[6] A. Chen, 'The Human Toll of Protecting the Internet from the Worst of Humanity', *The New Yorker*, 2017.

[7] A. Conn, AI Should Provide a Shared Benefit for as Many People as Possible, Future of Life Institute, January 2018.

[8] W. Min, Smart Policies for Harnessing AI, OECD-Forum, September 2018.

[9] AI Policy Challenges and Recommendations, The Future of Life Institute, undated.

This document is based on the STOA study on 'Ethics of artificial intelligence: Issues and initiatives' (PE 634.452) published in March 2020. The study was written by Eleanor Bird, Jasmin Fox-Skelly, Nicola Jenner, Ruth Larbey, Emma Weitkamp and Alan Winfield from the Science Communication Unit at the University of the West of England, at the request of the Panel for the Future or Science and Technology (STOA) and managed by the Scientific Foresight Unit within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament. STOA administrator responsible: Mihalis Kritikos.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

stoa@ep.europa.eu (contact)

http://www.europarl.europa.eu/stoa/ (STOA website)

www.europarl.europa.eu/thinktank (internet)

http://epthinktank.eu (blog)