EUI FLORENCE SCHOOL OF REGULATION

ROBERT SCHUMAN CENTRE

POLICY BRIEF

Unlocking the potential of AI: Opportunities and challenges for European policy [1]

On 14 January 2021, members of the European Parliament's (EP) committee on Artificial Intelligence in a Digital Age (AIDA) consulted with experts from the European University Institute (EUI) on topics in the regulation of artificial intelligence (AI).

After an introduction in which AIDA committee members outlined their goals and expectations for the role that AI will play in society, the EUI experts provided some general remarks on AI and algorithms, and then presented their original research. Their presentations covered the difficulties of writing laws and regulation for new technologies; economic interactions with AI and algorithmic collusion; and algorithmic content filtering for online platforms. The webinar concluded with statements and questions from the parliamentary groups.

In the view of the AIDA committee AI is a technology of strategic relevance. We see vast innovation taking place in this area, but unfortunately most of the action happens outside of Europe. To become more innovative and competitive, both more investment and more data to train algorithms are needed, and appropriate conditions must be established. The AIDA committee is also focused on developing a strategy for Europe to survive in a digital world and eventually assume a leadership role in AI.

[1] Special Committee on Artificial Intelligence in a Digital Age (AIDA) of the European Parliament Report on webinar with EUI experts, 14 January 2021, "Current state of play in AI research and applications"

Issue 2021/16
May 2021

Authors

Philip Hanspach, PhD Researcher, EUI
Marina Sanchez del Villar, PhD Researcher, EUI
The authors are members of the EUI's Interdisciplinary Cluster on Technological Change and Society

This policy brief draws from the experts' presentations and the questions the Members of the European Parliament (MEP) asked them. It expands on the topics that most interested the MEP and provides further references for policymakers interested in regulating this emerging technology.

Introductory remarks on AI

AI can "perform functions that require intelligence when performed by people".[2] It provides opportunities for individuals and society as a whole – with regard to sustainability, health and the possibility of increasing knowledge – but it also poses significant risks, such as unemployment, discrimination and social exclusion.

Academics distinguish between two types of AI:

1. Narrow, or specific, AI characterises systems capable of carrying out single specific tasks that usually require intelligence at a satisfactory level. This is the type of AI that is already with us, in many forms and applications. Famous examples include the algorithms trained with reinforcement learning to solve narrow challenges such as playing Chess and Go.
2. Strong, or general, AI comprises systems that exhibit most of the range of human cognitive skills, possibly at a superhuman level. To date, no one has succeeded in the development of general AI.

The current boom of AI is due to the recent transition from the knowledge-based approach to the use of machine learning (ML) algorithms. In the knowledge-based approach humans are both users and creators of the system's knowledge base. By contrast, ML algorithms learn from examples and correlations. AI algorithms can make predictions, and some can even take decisions accordingly, with or without human supervision.

Algorithms in society and the economy

Two issues are of central importance for the role of algorithms in society and the economy: cooperation between humans and algorithms, and algorithmic learning. At the heart of these issues lie challenges in economic organisations that predate the deployment of algorithms in the economy, allowing economists to use insights from game theory and industrial organisation to understand them and propose solutions.

1. Cooperation between humans and algorithms is a natural extension of challenges humans face in the workplace. The ability to cooperate effectively with co-workers, clients and other stakeholders is a vital skill, also in technical professions. Human cooperation relies on trust, effective communication and shared goals. These attributes also determine how algorithms can fit into society and how effectively they will work with humans. The notion of humans interacting with AI is profoundly novel and researchers at the frontier of human knowledge explore issues that arise here.

Computer scientists and economists see challenges in human–AI interaction both from a game theoretical perspective, which stresses rational choice, and from a behavioural perspective, which considers that individuals may not always behave rationally. An example for the game theoretic perspective arises in the programming of an autonomous car. Instructing the car to drive in a very defensive way and to always give way to avoid accidents may seem like a good idea. However, humans who anticipate such behaviour might try to take advantage of it and no longer obey the traffic rules when they encounter an autonomous car – ignoring a red light or taking the right of way because they expect the car to yield. This is problematic because it might encourage irresponsible behaviour, provoke dangerous traffic situations and hinder the effectiveness of, and acceptance for, autonomous cars, a technology with the potential to increase road safety.[3]

Other challenges are more behavioural in nature and cannot be addressed in the rational choice framework that underlies game theory. For example, humans have been found to be less cooperative when they believe that they are facing a machine. People's willingness to cooperate is influenced more by their beliefs than whether they are really interacting, or not, with a machine. Indeed, some experiments show a breakdown in cooperation when humans believe they are interacting with an algorithm – when in reality they are interacting with other humans. This explains why Google did not initially reveal that its chatbot Duplex was

[2] Kurzweil, R. (1990). The age of intelligent machines (Vol. 579). Cambridge: MIT press.

[3] Further information on this topic and on human-machine communication can be found in a report on a recent event at the EUI.

in fact a machine. The evidence shows that, sometimes, concealing the underlying algorithmic nature of a service may facilitate cooperation with humans. This is an important point for policymakers, who should realise this intrinsic trade-off between transparency and efficiency of algorithmic services.

2. Algorithmic learning poses a different policy challenge, either because algorithms learn to do things too well, which is the focus of this section, or not well enough, and in particular exhibiting problematic biases, the topic of the next section. Narrow AI astonishes researchers with superhuman capabilities and sometimes with its innovative strategies. Unintended consequences and problems can arise, however, exactly because algorithms may learn how to navigate precisely defined challenges extremely well. Two examples of such consequences include algorithmic collusion and outcomes of recommender systems.

Algorithms are used to update prices for goods quickly and flexibly both online and offline. For example, algorithmic pricing is common in commodities markets and it is also used to price fuel at gas stations.[4] Pricing algorithms have the potential to increase market efficiency by responding to changing market conditions, which can lead to beneficial deals for both firms and consumers. Algorithmic collusion refers to the risk that price-setting algorithms deployed by competing firms may tacitly agree to charge high prices. This is in principle also possible when humans set prices, but in practice difficult to sustain without explicit communication, a *per se* violation of competition law. For competing firms, it is typically profitable to undercut their rival at least in the short run, making it difficult to charge consistently high prices. Recent research in algorithmic collusion, however, has shown that algorithms can learn to charge consistently high prices over a long period of time.[5] Algorithms in this experiment even learned to recover from a temporary breakdown in collusion. They forgave mistakes and market perturbations instead of starting price wars – which can be costly for firms but are often beneficial to consumers.

A recommender system is a different kind of learning algorithm. This kind of technology is often deployed by firms operating digital platforms to recommend content to consumers, for example movies on a streaming platform, songs on a music platform or products on a retail platform. Recommender systems learn about the preferences of any consumer from her own and others' past observed behaviour. They can help make markets more efficient by drawing consumer attention to content they might not have found otherwise. However, this can be a double-edged sword, as recommender systems may give platforms great power to make certain products more visible. Products that are not recommended may disappear from the market. This is not problematic *per se* if product quality and consumer appeal determine a product's success. However, if platforms have a lot of market power, or even happen to be in a dominant position, there are concerns that recommender systems have a large and non-transparent influence on product success. The abuse of a dominant position might materialise when platforms use their recommender systems to preferentially promote their own products. Without proper algorithmic regulation, this practice might be hard to monitor or even discover.[6]

Algorithmic bias and its sources

Academic experts and policymakers are also concerned about algorithmic bias. In many cases, the bias comes from the data itself, as the algorithm learns patterns in the data that humans may not perceive, and then provides predictions based on these biased data. Economists study the effect of such biased predictions on economic outcomes and have identified different sources of biased algorithmic predictions:[7]

1. **Unrepresentative training samples**. Algorithms are first trained on a dataset before they are tested. This means that the data used to train the algorithm is of utmost importance.

[4] Assad, S., Clark, R., Ershov, D., & Xu, L. (2020). Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. *SSRN working paper.*

[5] Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., & Pastorello, S. (2020). Protecting consumers from collusive prices due to AI. *Science*, *370*(6520), 1040-1042. https://cadmus.eui.eu/bitstream/handle/1814/69109/CALZOLARI_ET_AL_2020.pdf?sequence=1.

[6] The abuse of a dominant position is prohibited by EU law under Article 102 of the Treaty of the Functioning of the European Union.

[7] Cowgill, B., & Tucker, C. E. (2020). Algorithmic fairness and economics. *The Journal of Economic Perspectives*. https://dx.doi.org/10.2139/ssrn.3361280

Existing biases in the historical data may persist in the algorithm and would most likely lead to biased predictions when deployed in a broader population. An example is a bank that is interested in predicting loan performance. The algorithm will be trained on the bank's loan portfolio, which is a highly selected set of loans and may not include characteristics of loans for projects that were rejected by the bank.

2. **Mislabelling outcomes in training samples.** Sometimes the bias might be hard-coded in the data. One can think of a training database that contains variables that are subjective, such as likeability or measures for soft skills. The algorithm would take these variables and associate individuals with certain characteristics to lower levels of likeability. Mislabelling can also arise from measurement issues. If the databases include only variables that are easily measurable, and associate a label to them denoting an outcome that should also encompass non-measurable variables, the algorithm will optimise its predictions based exclusively on measurable characteristics. An example of this is job performance, which is composed of quantifiable as well as more abstract and hence harder-to-measure elements.

3. **Biased programmers.** Although there is no empirical evidence of programmers consciously building biased algorithms, there are numerous examples of algorithms built by a homogenous group of programmers that provided biased predictions when released into society. A lack of diversity among programmers may make them less perceptive of unrepresentative training samples or mislabelling of outcomes. This is a key issue, as developers are regularly the ones performing most of the tasks: from gathering the data to writing the algorithms. Anecdotal evidence of this bias includes Amazon's recruiting algorithm (trained on Amazon's existing labour force, it tended to favour men over women) and Google's image recognition (trained mainly on white subjects, it would sometimes mislabel black people as animals[8]). There are several initiatives to increase diversity in the programming workforce and more generally among STEM majors (science, technology, engineering and mathematics). In particular, members of the AIDA committee called for more inclusion of women in STEM education and in panels when discussing AI.

4. **Algorithmic feedback loops.** Algorithmic feedback loops are a concern over the long term. There are certain instances when the predictions of the algorithms are used to make decisions that affect individuals. The outcome, in turn, is used in the data on which subsequently other algorithms will be trained. There is empirical evidence of this feedback loop in the case of recidivism and bail granting.[9]

Upload filters and freedom of speech

Online platforms that allow users to connect and interact are an important part of the Internet. Users publish their own content on platforms such as YouTube, Twitter, Facebook or LinkedIn to inform others, express themselves or engage in discussion. However, there is a justified concern about harmful behavior: aggressive and uncivil discussions, misinformation and political polarisation, as well as spam, trolling, and fraud. A recent report assessing the impact of algorithms for online content filtering argues that algorithms that are used to automate moderation, so-called upload filters, are indeed "needed to monitor the huge amount of material that is uploaded online and detect (potentially) unlawful and abusive content".[10]

However, there are many challenges to identifying and filtering out inappropriate content. The commonly used systems can only predict the probability that any piece of content is harmful, so mistakes are unavoidable. A dilemma arises because stricter filtering will remove a greater amount of harmful content but will also censor more innocuous content. The possibility of such errors spurs fears that the use of content filters may limit freedom of speech, especially if filters make systematic errors and exhibit bias towards content, e.g. of a particular political platform.

[8] The Verge (January 12, 2018). Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.

[9] Cowgill, B. (2019). The impact of algorithms on judicial discretion: Evidence from regression discontinuities. Technical Report. Working paper. http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf.

[10] Giovanni Sartor & Andrea Loreggia (2020). The impact of algorithms for online content filtering or moderation - Upload filters. https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)657101.

A fundamental problem for all the main filtering techniques[11] is that the 'ground truth' of what is considered harmful is ultimately a human decision. Therefore, any potential threat to freedom of speech is a human issue, not a technological one. Policymakers need to acknowledge the constraints that platforms face. First, they need to understand that content filtering cannot be perfect. Second, the filtering rules are ultimately governed by the humans that formulate them. Third, content filters can only be judged fairly against the relevant counterfactual – that is, manual checking by human operators. Finally, the moderating technology employed needs to be able to handle the vast amount of content platforms face. The use of automated filters is therefore necessary, especially for small- and medium-sized companies, if they want to compete with better funded companies. This suggests policies aimed at creating more legal certainty around the use of upload filters, increased transparency and improved appeal mechanisms for users, as well as support for small-and medium-sized enterprises to automate platform moderation.

Strategic global acquisition and provision of data

As noted by the AIDA committee members, the fuel on which any AI tool runs is data. Therefore, provision of large amounts of high-quality data must be a key part of the EU's strategy.

One of the consequences of this reliance on data is that the companies that accumulate the most data can have a strong advantage, which eventually creates a virtuous cycle: more data gives better predictions, which are used to increase sales and attract more customers, customers who in turn provide the company with more data, and so on. This cycle, which is not necessarily a negative outcome (with recommender algorithms, for example, customers get more targeted products, which can lead to an increase in consumer surplus) could eventually tip markets and generate winner-takes-all situations.

Because access to and accumulation of data can also lead to good market outcomes, it follows that a blanket restriction on data collection is not a good solution. The key element that needs to be considered is keeping the competition in these markets open, rather than restricting the amount of data collected, which might deteriorate the efficiency of algorithms.

This pertains also to data generated or owned in different jurisdictions. Cross-border data flows have recently attracted growing interest among policymakers. In the last decade, different countries have increased their restrictions on the cross-border flow of data. Some of the factors that contribute to this trend are privacy concerns, data sovereignty, cybersecurity, and industrial protectionism.

A recent 2021 World Bank report[12] provides a taxonomy of these different practices, ordered below from least to most restrictive:

- No restrictions for cross-border data transfers;
- Obligation for local storage: the data must be stored in the home country but might be processed and transferred elsewhere;
- Obligation for storage and processing: the data must be stored and processed in the home country but might be transferred elsewhere;
- Conditional flow regime: a hybrid approach by which the home country imposes some conditions to allow the transfer of data;
- Total ban on transfers: no cross-border transfer of any copy of the data is allowed.

One can draw the following conclusions from the report: the EU should make sure that European companies face the most favourable terms possible when operating with European data. By leveraging on the European community, companies can have access to a richer pool of data, which can make their AI tools competitive on the international scene.

[11] These techniques include metadata searching, blacklisting, and artificial intelligence-based solutions.

[12] Ferracane, M. F. & van der Marel, E. (2021). Regulating Personal Data: Data Models and Digital Services Trade. Policy research paper 9596, World Bank Group, http://hdl.handle.net/10986/35308.

The need for a strategic data policy on the international stage also follows from the initial concern of policymakers for the EU's standing relative to foreign countries. The report implies that the ability of the EU to continue shaping AI policy is closely linked with the standing of its companies in the development and use of AI technology. If the EU wants to regulate AI applications of Chinese or US origin, it will be difficult to argue that burdensome regulation is not in fact a hidden barrier to trade. If EU policy is perceived as being out-of-step with American and Chinese regulators and disproportionately targeting US and Chinese companies, this might introduce legal risk and potentially result in retaliation by non-European countries. If European companies do not successfully develop and use AI, the EU's potential to shape AI policy will diminish over time.

Options for regulating AI well

Policymakers are currently considering a variety of measures to counter the problems posed by AI.[13] Among these measures is algorithmic certification, a practice normally discussed in relation to price-setting algorithms, although it could be applied in other settings as well. The certification of algorithms, however, is very complex to carry out. AI tools evolve at a fast pace and it would be hard for any regulator to keep up with the technology. This certification solution is likely to decrease innovation and would also deprive economic actors of the potential benefits of AI applications in different sectors. Instead, experts recommend focusing on algorithmic auditing.[14]

Auditing is tightly linked to transparency and explainable AI (XAI). Auditing requires that the companies using AI can identify the input variables that determine their results. Practices that promote algorithmic auditing allow the users to inspect the algorithm's predictions and verify how it uses the data it receives. The auditing could also be done by a governmental agency, which could check that the predictions of the algorithms do not lead to collusive pricing or biased decisions, for example. This way, supervisors could modify the value of the inputs and verify how the predictions change. Notice how, in a sense, auditing has the potential to correct biases, because algorithms are more auditable than humans.

In addition, regulators in some sectors are joining the private sector in the use of AI tools. For example, in the banking sector, the so-called Reg-Tech uses AI to monitor the complex activities carried out by banks. Although this is still an emerging trend, it is a reminder that that the public sector, as well as the private sector, can employ AI and ML tools to perform its functions.

How should we think about regulating these activities? Counterfactual thinking is important to put risks in perspective and allow a reality-focused cost–benefit analysis. Where AI can replace humans, we can judge the benefit from using AI by comparing the performance of AI with those of humans. Failing to do so and judging AI by any other standard, for example a zero-tolerance policy for errors, might lead to consumers missing out on the benefit that AI can bring. For example, we should judge the performance of self-driving cars against the performance of human drivers. If the adoption of self-driving cars results in fewer accidents, they should be adopted even if this does not fully eliminate the possibility of traffic accidents. In this example, even if the advantages of self-driving cars over human drivers were only small, a ban on them would eliminate the ability and incentive for developers to further improve self-driving cars. Failing to measure the outcomes of AI deployment against the relevant counterfactual may therefore lead to policies that are too restrictive. The consequences are harm to citizens by banning potentially beneficial technologies and preventing the improvement of such technology.

As for the regulation of AI technology itself, we need to distinguish between two exercises, which require two different intervention levels.

1. To keep up with the technology, policymakers need as much information as possible from different stakeholders. This requires the ability to collect relevant and timely information, which is better achieved with pluralistic and decentralised institutions.

2. Investment in AI innovation requires European-level intervention. Since the single market is incomplete, there remain regulatory, cultural and economic obstacles for the learning and scaling of national systems to develop on par with US and Chinese technological capabili-

[13] European Commission (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[14] Seele, P., Dierksmeier, C., Hofstetter, R., & Schultz, M. D. (2019). Mapping the ethicality of algorithmic pricing: A review of dynamic and personalised pricing. *Journal of Business Ethics*, 1-23. https://doi.org/10.1007/s10551-019-04371-w.

ties. The only way to tackle this disadvantage is actions, on the European level.

In more practical terms, there are voices calling for a risk-based approach to the regulation of algorithms.[15] In this approach, each algorithm (or algorithm-application pair) would be classified into a risk category, and each category would have to meet certain requirements to be allowed to operate. Although the *status quo* of this type of regulation is a binary, high/low classification, some stakeholders are advocating a finer gradation that goes beyond this binary set-up.

Finally, the realm of AI technologies and tools is wide and is rapidly changing. Regulation practice must remain flexible and able to adapt to the pace of the technology. The use of regulatory sandboxes can help in this regard. Sandboxes consist of circumscribed areas of the territory, the population or a company's activities, where decision-makers loosen the regulation so that companies can experiment with new tools. The advantage of running these experiments is that inside the contained environments there are sufficient controls to avert catastrophic risks. The companies benefit because they do not need to meet all prior legal requirements (procedural or substantive) and can test new approaches (in this case, new algorithms). The benefit for the regulators lies in their capacity to test their legislation before implementing it in society as a whole.

Conclusion

The large-scale deployment of AI and improvements in existing algorithms in European firms hold opportunities for European citizens, who face algorithms in their roles as consumers as well as employees, entrepreneurs and firm owners. Some of the opportunities include better cooperation between humans and machines, for example in the workplace or on the road; expert systems with superhuman abilities in narrowly defined tasks; more efficient markets through more flexible pricing; better recommendations to consumers on platforms with large offers; and less harmful content on online platforms.

At the same time, many of these areas harbour the risk of negative outcomes if important policy challenges are ignored: cooperation might break down if humans do not trust the machines they work with; pricing algorithms may coordinate to individually set high prices; recommender systems can increase the market power of firms operating large platforms; algorithms may exacerbate biases already present in society; and automated content moderation might block innocuous content and limit freedom of speech.

To enjoy the potential benefits of AI while managing the risks, policymakers need to adopt an appropriate risk framework that balances the potential outcomes from adopting, regulating or banning different AI applications. The task is a knife's edge, as it requires taming the negative consequences while making sure that the positive aspects of AI are not curtailed. A sophisticated risk framework should not just consider the immediate impact of any technology, but also consider the relevant counterfactual and the larger-scale risk of failing to strengthen AI-driven European firms.

In this report, we have outlined and expanded on the main concerns that members of the European Parliament expressed to EUI experts in the AIDA committee meeting held online on 14 January 2021. The exchange showed the value of continued consultation between policymakers and academia. The committee will continue to follow closely topical problems in the regulation of new technologies.

[15] Data Ethics Commission (2020). Opinion of the Data Ethics Commission. https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html.

The Florence School of Regulation

The Florence School of Regulation (FSR) was founded in 2004 as a partnership between the Council of the European Energy Regulators (CEER) and the European University Institute (EUI), and it works closely with the European Commission. The Florence School of Regulation, dealing with the main network industries, has developed a strong core of general regulatory topics and concepts as well as inter-sectoral discussion of regulatory practices and policies.
Complete information on our activities can be found online at: fsr.eui.eu

Robert Schuman Centre for Advanced Studies

The Robert Schuman Centre for Advanced Studies (RSCAS), created in 1992 and directed by Professor Brigid Laffan, aims to develop inter-disciplinary and comparative research on the major issues facing the process of European integration, European societies and Europe's place in 21st century global politics. The Centre is home to a large post-doctoral programme and hosts major research programmes, projects and data sets, in addition to a range of working groups and ad hoc initiatives. The research agenda is organised around a set of core themes and is continuously evolving, reflecting the changing agenda of European integration, the expanding membership of the European Union, developments in Europe's neighbourhood and the wider world.

www.eui/rsc

Co-funded by the
Erasmus+ Programme
of the European Union

Views expressed in this publication reflect the opinion of individual authors and not those of the European University Institute or the European Commission.

doi:10.2870/50768
ISBN:978-92-9084-971-1
ISSN:2467-4540
QM-AX-21-016-EN-N