BRIEFING
STOA Options Brief

European Parliament

Auditing the quality of datasets used in algorithmic decision-making systems

The existence of biases pre-dates the creation of artificial intelligence (AI) tools. All human societies are biased – AI only reproduces what we are. Therefore, opposing this technology for this reason would simply hide discrimination and not prevent it. It is up to human supervision to use all available means – which are many – to mitigate its biases. It is likely that at some point in the future, recommendations made by an AI mechanism will contain less bias than those made by human beings. Unlike humans, AI can be reviewed and its flaws corrected on a consistent basis. Ultimately, AI could eventually serve to build fairer, less biased societies.

1. Main findings

Based on what has been presented in the study, the following conclusions can be reached:

1. The fight against biases is not specific to datasets or AI, as a human operator can introduce biases that are much more accentuated and more difficult to eradicate. Therefore, any criticism of the biases derived from the use of AI systems must contemplate that their alternative – the human element – may incorporate the same, or worse, biases.
2. Biases are inherent to human beings, their culture and their history. There are multiple taxonomies of bias, and identifying and classifying them all is complex. AI-based solutions incorporate new biases and tend to magnify existing human biases. To identify and mitigate bias in AI-based solutions, it is necessary to understand and be aware that biases are introduced at all stages of the process of AI development by the training dataset, the algorithm, and the humans involved.
3. An essential step to mitigate biases is to create or use high quality domain-specific training datasets to guarantee a fair knowledge representation of the 'real-world' to the AI-based system. Mechanisms of oversight and accountability should be implemented to continuously assess the quality and integrity of the data.
4. Techniques exist that correct biases in AI systems via pre-processing, in-processing, or post-processing, to achieve greater fairness in those systems. Currently, several companies are developing toolkits to help in this process, although there is still a lack of standardisation in the sector.
5. Best practices recommend including the 'human in the loop' during the development process, and building diverse, interdisciplinary development teams with ethical reflection and inclusive participation.
6. It is not easy to design a regulatory framework that is able to deal with bias, since bias is a complex concept that is not synonymous with discrimination, at least from a legal point of view. In the EU context, the prohibition of discrimination is limited to particular contexts and concrete factors. This scenario introduces doubts about allowing the use of algorithms that introduce bias in some specific cases, and needs to be addressed as soon as possible. There are different ways to deal with this situation. For instance, incongruences and contradictions could be eliminated by improving the current regulations, along the lines contained in the proposal for a Council directive on implementing the principle of equal treatment between

persons irrespective of religion or belief, disability, age or sexual orientation. However, alternative legal tools able to cope with discrimination issues could also be considered.

7. The General Data Protection Regulation (GDPR) could provide an excellent tool to fight against bias through its concept of 'fairness'. However, it also shows some weaknesses. If the main problem of anti-discrimination law is its limited application to the multiple forms of unfair treatment produced by algorithmic systems, the main problem of data protection law is its lack of coverage of databases that do not contain personal data. Anonymisation techniques are appropriate to protect privacy, however, they do not protect against the reproduction of biases when using such data. This is aggravated by the fact that the GDPR principles are not applicable to anonymised data.
8. The creation of standards and certificates applicable to datasets and AI mechanisms is a fundamental pillar in the regulation of these assets. However, they are each in their early stages, both in datasets and in the AI arena. Standards related to datasets should include information about the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty. On the other hand, standardisation and certifications related to datasets and AI tools must allow for flexibility in order to be able to include the variety of possible data formats and collections used in AI applications.
9. Monitoring of high-risk AI tools is certainly needed if we want to mitigate bias. However, monitoring might become extremely complex, for several reasons. This means that we must create adequate tools able to deal with such complexity. Dynamic monitoring must be carefully considered, as the draft AI act proposes. The governance framework introduced by the AI act is definitely complex. It includes supervision by third parties, with competences shared by the Member States and the European Commission as last resort to ensure compliance. However, it does not provide individual citizens and non-governmental organisations (NGOs) protecting human rights with adequate tools to complain to market surveillance authorities or to sue a provider or user for failure to comply with the requirements.

2. Policy options and assessment

Some policy options that may serve to improve the current situation are suggested in this section. Some of them are complementary and others are mutually exclusive. In this last case, reasons for supporting one or the other are provided, so that policy-makers can gain a better understanding of the present situation. More detailed information on the proposed policy options listed here can be found in the full version of the study.

2.1 No new legislation focusing on biases is required

As an alternative to creating new legislation, the EU may rather focus on tackling the problem of misalignment between the different regulatory tools. A first and tempting option to address the issue of bias is to develop specific legislation on this matter. If this were done at the EU level through legislation, such as a directive or a regulation, we would achieve greater unification of criteria among all Member States when responding to this complex issue. We would also gain in legal certainty and specificity of the applicable legal framework. However, it is not currently a good idea to introduce a specific regulation on biases in datasets or on biases as a result of the use of AI mechanisms in general. All this, in short, makes it much more advisable not to produce a specific regulation, but to take advantage of the opportunities presented by that which already exists. In any case, slight modifications to the regulations that are now in the approval phase would be advisable. This is particularly important in order to promote best practice, such as the use of standards and the implementation of certification systems, and to solve the misalignment between the different regulatory tools (in particular, the GDPR and the new regulations), which may generate legal uncertainties for all bodies concerned. Otherwise, companies may find themselves in breach of regulations at the intersection of the AI act, GDPR and the general product safety regulation.

2.2 Strengthening bias mitigation in the data collection

The application of best practices for bias mitigation is a preventive approach that should be a policy priority from the start of data collection onwards. This means that data collection should comply with the FAIR principles (FAIR stands for: findable, accessible, interoperable and re-usable). The fact that FAIR principles are respected at the time of data collection will also serve as an extremely efficient measure both for introducing standards for better data governance and for auditing the databases in which they are integrated. In addition, a proper implementation of FAIR principles would allow a much more efficient integration of separately constructed databases, so that a better control of biases at their source can be ensured. This means, of course, including adequate information about the characteristics of the data they contain, namely: the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists, which should be described in a publicly available and consistent manner.

2.3 Promote database certification

Certification could enforce bias-awareness from the very beginning of the lifecycle of algorithmic systems. This policy option explores the possibility of introducing certificates for datasets with the objective of avoiding biases. This comprises two main possible actions: first, certifying that the dataset developer has applied the best available practices to avoid the presence of significant biases; second, certifying that the dataset developer provides accurate relevant information about the dataset that may prevent other stakeholders from developing or using the dataset and the algorithmic system in a biased way. Of course, these certificates on the datasets could be complemented by others referring to the artificial tool that has been shaped thanks to these data. Some options advocate certificates obtained exclusively from public entities, while others allow self-certification. It is also possible to introduce different requirements depending on the level of risk involved in a given treatment.

1. Mandatory certification for databases that will feed high-risk AI systems: making data providers accountable;
2. Voluntary certification for high-risk AI system databases: a tool to comply and demonstrate compliance for AI system providers.

2.4 Granting transparency rights to AI-system-subjects

This option could open a window to finding the source of biased results. The rights to access and information provided by the GDPR include 'meaningful information about the logic involved' in automated decision-making and profiling. However, this 'meaningful information' does not extend to information about training datasets that are key to the logic involved in the automated processing. Under this policy option, AI-system-subjects should be able to ask for meaningful information about training datasets. This information should, at least, include the metadata (or datasheets) about training datasets generated with certification that could serve this purpose. The information should not adversely affect the rights or freedoms of others, including data protection rights, trade secrets or intellectual property. However, the result of this balance should not be a refusal to provide relevant information to the AI-system-subjects.

Unfortunately, individual rights for AI-system-subjects are not currently included in the proposed artificial intelligence act. This should be changed if we do not want to risk that the rules on prohibited and high-risk practices become ineffective in practice. For this purpose, it would be necessary to ensure that such information rights for decisions that significantly affect AI-system-subjects are also extended to hybrid decisions, or recommender and decision-support systems, and not only to decisions based solely on automated processing (contrary to Articles 13(2)(f), 14(2)(g) and 15(1)(h) in the GDPR). It would not be advisable to implement this perspective of individual rights without other policy options that reinforce other ways of holding data and AI system providers accountable. The information provided by transparency rights alone is rather inadequate to govern algorithmic systems and their data flows.

2.5 Facilitating the implementation of the proposed AI act

Finally, it must be considered that compliance with the legislation currently under negotiation will entail costs and risks for companies. This might be challenging, especially for small and medium-sized enterprises (SMEs), or companies operating in international markets. Were this the case, their competitiveness would probably suffer severe damage. Indeed, the impact assessment accompanying the proposed AI act acknowledged that those SMEs that supply high-risk AI systems would, in principle, be more affected than large companies. Thus, if the new regulations are aimed at boosting the European Union's technological and industrial capacity, as well as AI uptake across the economy, they should consider some measures able to make things easier for the industry and, in particular, for SMEs. AI regulatory sandboxes are excellent examples of such measures. Similarly, the obligation to consider SMEs' interests when setting fees related to conformity assessment will surely help them reduce their costs. Furthermore, the measures included in Article 55.1 of the draft AI act (such as providing SMEs with priority access to the AI regulatory sandboxes) are excellent steps in the right direction. However, these measures will not serve well to erase all additional costs caused by the new regulations. As the Impact Assessment acknowledges, ''whether the additional costs can at the margin discourage some SMEs from entering into certain markets for high-risk AI applications will depend on the competitive environment for the specific application and its technical specificities''.

Therefore, it is necessary to consider other initiatives that can be used to reduce the costs faced by EU companies, especially (but not exclusively) SMEs. Indeed, the draft AI act states that 'the framework will envisage specific measures supporting innovation, including regulatory sandboxes and specific measures supporting small-scale users and providers of high-risk AI systems to comply with the new rules'. In our opinion, one of the most appropriate measures would be for public institutions to make high-quality databases available to private agents. This would substantially reduce the expenditure associated with the review of these datasets and the corresponding certifications. These initiatives could, of course, be complemented by specific subsidies aimed at helping companies to adapt to the new regulatory framework. The most appropriate approach would probably be a type of aid that would alleviate their substantial one-off cost for market entry.

This document is based on the STOA study 'Auditing the quality of datasets used in algorithmic decision-making systems'. The study was written by Iñigo de Miguel Beriain, Pilar Nicolás Jiménez (UPV/EHU), María José Rementería, Davide Cirillo, Atia Cortés, Diego Saby (Barcelona Supercomputing Center), and Guillermo Lazcoz Moratinos (CIBERER - ISCIII) at the request of the Panel for the Future of Science and Technology (STOA), and managed by the Scientific Foresight Unit, within the Directorate General for Parliamentary Research Services (EPRS), European Parliament. STOA administrator supervising the study: Andrés García Higuera. STOA administrator responsible: Philip Nicholas Boucher.

DISCLAIMER AND COPYRIGHT

stoa@ep.europa.eu (contact)

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

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

http://epthinktank.eu (blog)