Special Committee on Artificial Intelligence in a Digital Age

European Parliament

AIDA Working Paper on
AI and Bias

following the AIDA Public Hearing on 30 November 2021

November 2021 | WWW.EUROPARL.EUROPA.EU - @EP_ARTIFINTEL

Introduction

The special committee on Artificial Intelligence in a Digital Age (AIDA) organised a public hearing on the topic of ' AI and Bias' on 30 November 2021. The hearing explored the impact of bias on the development of AI and strategies to reduce bias in AI systems. Featuring two expert panels representing diverse views from research, industry, and civil society, the event provided AIDA Members and the public with a variety of perspectives on the origins of bias in AI, the impact of bias on current and future uses of AI, and recommendations for strategies to combat such bias.

The AIDA Chair Dragoș Tudorache made his opening remarks by highlighting that the issue of bias plays a role in nearly every discussion about AI. In this light, uncovering the different source of biases and exploring their practical implications remains at the forefront of the AI discussion. "There has not been a single conversation we have had during AIDA's mandate - whether it was concerning competitiveness, democracy, health or defence - where the element of bias in algorithms has not been central to the debate. Bias has many sources and can arise in unexpected ways. For example, when AI developers have cognitive prejudices, or when datasets are skewed either historically or because they are built for a certain purpose. Our duty is not only to mitigate these biases and to prevent them from creeping from the 'real' world into the digital world, but also to leverage technology and AI to ultimately eliminate them from our societies," Mr Tudorache concluded.

Key Takeaways - Panel I:

The first panel focused on the impact of bias on the development of trustworthy AI and featured the following four panellists:

- **Timnit Gebru,** Independent Scholar
- **Victoria Espinel,** President and CEO, BSA | The Software Alliance
- **Michael O'Flaherty,** Director of the EU Agency for Fundamental Rights
- **Daniel Leufer,** Europe Policy Analyst, Access Now

Ms Gebru, an independent scholar and former employee of Google, made it clear that the number one objective of regulators to combat bias in AI should involve empowering tech workers and curbing the power of major tech corporations. Ms Gebru spoke of the importance of facilitating transparency and changing the incentives within Big Tech business models as the only way for companies to invest in technologies that benefit people and not merely seek to maximise profits. **Removing bias from AI involves creating checks and balances between tech companies and regulators, which is made possible primarily through empowering individual employees and whistle-blowers to speak out against the harms they experience from AI or company practices,** Ms Gebru stated. Ms Gebru provided an example of legislation that works to improve transparency and the protection of whistle-blowers, namely the Silenced

No More Act[1], which was adopted in the state of California in October 2021. The Act prevents the use of non-disclosure agreements by companies in cases of worker allegations of harassment, racism and other forms of bias and discrimination. Ms Gebru also spoke about the importance of diversifying the AI community and expanding the pool of AI research beyond that funded by Big Tech and the military, noting that "contrary to the Cold War rhetoric of the AI armed race, this is what really stifles innovation". Ms Gebru argued that **it is not tenable to have the same people setting the agenda on AI investment and research across Big Tech, academia, government, and the non-profit world, and that alternatives are urgently needed in the form of increased government investment into more diverse communities of AI researchers.** Ms Gebru also spoke about misinformation and hate speech on social media platforms, and how data from these platforms can feed into AI models such as large language models, with potentially devastating effects. While Facebook has recently been under increased scrutiny, other platforms including YouTube, TikTok, Telegram and Clubhouse are not receiving the same amount of attention.

Ms Espinel of the Business Software Alliance (BSA) highlighted that **bias can arise in all phases of the AI lifecycle which is why developing trustworthy AI must be seen as a comprehensive process taking place throughout the design, development, and deployment of AI technology.** Ms Espinel laid out the key elements of BSA's "Framework to Build Trust in AI"[2], which consists of impact assessments, risk mitigation practices, and organisational accountability. According to BSA, impact assessments should take into account the impact an algorithm has on people, the context and purpose of the system, the degree of human oversight, and the type of data used to train the system. Risk mitigation practices involve identifying the intended use and the potential harms of an AI system and should be conducted in the initial phases of development. Regarding organisational accountability, responsibility must be tailored to each companies' role in the development process of AI and must be able to distinguish between these. Ms. Espinel emphasised that it is important to remember that at every phase in the AI lifecycle, a broad range of stakeholders must take part in the process of creating and deploying AI to maintain transparency.

Mr O'Flaherty of the European Union Agency for Fundamental Rights made the point that **fostering awareness the implications of bias in AI requires careful oversight and an interdisciplinary approach.** Mr O'Flaherty put forth **six recommendations to combat bias** in AI:

1. Working with people affected by biased AI and fostering diverse workplaces in industry, in oversight, and in policy spaces
2. Dispelling the misconception that AI with no protected characteristics data (such as that relating to gender, race, or sexual orientation) cannot discriminate, and acknowledging research findings that show that general data can serve as a proxy for protected characteristics
3. Considering the legal complexity of bias in AI, remembering that bias can often only be perceived as discriminatory in specific contexts
4. Focusing on data quality, avoiding training data that promotes discriminatory behavior or results in underrepresentation of certain groups, and keeping a close eye on how feedback loops may promote bias
5. Requiring ex ante human rights assessments to check the quality of AI, in particular as regards high-risk AI applications and their potential impact on fundamental rights
6. Investing in further research on bias in AI, keeping in mind the interaction of technology, ethical specialists, and human rights experts in AI research

Mr Leufer of the human rights NGO Access Now spoke about the work of Access Now and partner NGOs in ensuring that fundamental rights remain at the forefront of the conversation around the development and regulation of AI, referencing the **2018 Toronto Declaration on Protecting the Right to Equality in Machine Learning**[3]. Mr Leufer proposed that the problem of bias in AI must be reframed to focus on more structural solutions and not simply around specific outcomes linked to bias. **A central focus in combating bias must be to recognise that the increased use of AI systems will lead to a general increase in surveillance and data collection practices in our societies, which often shifts the responsibility of guaranteeing human rights towards large tech companies and away from public institutions.** One recent example of this trend could be seen through the expanded use by universities and other educational institutions of proctoring software during the COVID pandemic for the purpose of remote exam monitoring, with documented discriminatory effects on students with darker skin. According to Mr Leufer, the focus on bias may be premature in this case because it overlooks the greater question of whether or not these types of surveillance applications even have a place in a society that values privacy and human rights. Mr Leufer also shared recommendations on the draft AI Act[4], arguing that the proposed draft Article 5 on prohibited AI practices should be made easily amendable in future to facilitate updates in view of new problematic AI use cases that may emerge over time. In this regard, Mr Leufer referenced the **2021 Civil Society Statement - 'An EU Artificial Intelligence Act for Fundamental Rights'**[5], developed by Access Now and 110 other NGOs.

Key Takeaways - Panel II:

The second panel featured three speakers who discussed reducing bias in AI systems through algorithmic accountability and data governance.

- **Nathalie Nevejans,** AI & Robotics Law and Ethics Expert
- **Dana Rao,** Executive Vice President, General Counsel and Corporate Secretary, Adobe
- **Ricardo Baeza-Yates,** Director of Research, Institute for Experiential AI, Northeastern University, USA

According to Ms Nevejans of the University d'Artois, self-regulation alone is not sufficient in ensuring the development and deployment of unbiased AI systems because the task of making AI models fair and bias-free is often a difficult and expensive exercise for businesses to undertake. Therefore, it is important for the EU to regulate and achieve legal certainty in this area, notably through its proposed AI Act. Ms Nevejans highlighted the importance of clarifying the definition of bias in the draft legislation, because of the need to distinguish between different notions of bias, such as unfair bias, which should be ruled out as a matter of principle, and possibly acceptable biases, which may be necessary or even desirable in certain AI applications. Ms Nevejans also raised the question of whether the proposed risk-based approach of the AI Act may ultimately send the wrong message to AI businesses and the wider AI community, namely that only high-risk AI applications should be screened for bias whereas many low-risk AI applications may also result in harmful effects based on bias. In this regard, Ms Nevejans recommended a possible reassessment of the AI risk categories proposed under the AI Act, citing the example of chat-bots, which are currently only classified as low risk despite the large power they have to spread misinformation online and the fact that they are often trained on biased data sets. More broadly, Ms Nevajans recalled that **building trust in the use of AI is essential to combatting bias, which is possible only when algorithmic decisions respect rights and values, have explainable outcomes, and provide remedies against the harms of AI assisted decisions.**

Mr Rao of Adobe highlighted that in seeking to reduce bias in AI, legislators and developers alike should focus on decreasing harmful outcomes rather than aiming to remove bias entirely. Mr Rao recalled that the value of AI comes from learning about tendencies and patterns in data and using these insights to solve specific problems. **Bias may actually be a desired outcome in some AI applications, such as online marketing where a company is targeting a specific demographic.** For example, when selling women's products, the desired outcome of an algorithm would be one that favours women. According to Mr Rao, there are many such instances where one may not want to remove bias entirely or where bias may even play a desired role. Furthermore, Mr Rao stated that privacy laws ought to be addressed with the reasoning that policy makers must enable innovators to obtain clean, **privacy-friendly data, which assists in the creation of fair AI.**

Lastly, Mr Baeza-Yates of Northeastern University gave an overview of sources of bias in AI, emphasising that bias is present not only in data, but also in other aspects of AI systems, such as the optimisation function used and the feedback loop between the system and its users. Mr Baeza-Yates underlined that **societal problems like racism and sexism cannot be solved by technology alone and that technology amplifies biases that are already existent in society.** As a result, bias is a problem that must be primarily addressed on a societal level, but the impact of technology must not be underestimated, and tech companies have a responsibility to reduce bias. In this regard, Mr Baeza-Yates argued that **voluntary risk assessments and codes of conduct are not sufficient to reduce bias in AI because these present a conflict of interest for companies and lead to frequent scenarios where "monetization trumps ethics".** Due to this conflict of interest, Mr Baeza-Yates believes regulation is necessary however stressing the point that regulation should not be directed at a specific type of technology but rather at solving a particular problem. As regards the EU's draft AI Act, Mr Baeza-Yates questioned the validity of its risk-based approach, as **risk in AI can be seen as a "continuous variable", and businesses may attempt to downplay the risks of their AI systems** so as to evade regulation. Mr Baeza-Yates also recalled that bias in AI is not only a problem in the oft-discussed social media context, but also in applications linked to e-commerce, **where exposure bias and popularity bias**, among other issues, should be addressed by policy-makers.

1 SB-331 Settlement and Nondisparagement Agreements (2021-2022) https://bit.ly/3dw5FtY

2 Confronting Bias: BSA's Framework to Build Trust in AI, The Software Alliance https://bit.ly/30eRal0

3 The Toronto Declaration, https://www.torontodeclaration.org/

4 Proposal for a Regulation Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts (COM/2021/206 Final)

5 'An EU Artificial Intelligence Act for Fundamental Rights - A Civil Society Statement', https://www.accessnow.org/eu-artificial-intelligence-act-fundamental-rights/

Contributions by Political Groups

Disclaimer:
The statements herein below are drafted by the political groups of the European Parliament. The opinions expressed are the sole responsibility of the author(s) and do not necessarily represent the official position of the European Parliament. This document may contain links to websites that are created and maintained by other organisations. The AIDA Committee and the AIDA Secretariat do not endorse, nor are they affiliated with, the view(s) expressed in the said statements nor through the said websites.

The European People's Party Group (EPP)

The many cases of bias in AI are worrying and reducing bias in AI is a quite crucial topic. Issues of bias in technology are not new and we already know they are mainly based on negligence and prejudice. Quality control and standards prevent rushed products, sloppiness and negligence. Tough laws, social debate and the power of a transparent market - where customers can choose - limit hatred or prejudice.

When referring to bias in AI, we are talking about systematically producing unfair outcomes for specific demographic groups and it is important to underline that there is no universal definition of fairness. Consequently, the complexity of this question is also linked with the choice of the best metric and method for mitigating potential risks.

The EPP Group believes that we can mitigate bias in AI by developing systems to test and validate AI systems that are able to detect and monitor bias, and by ensuring a strong societal debate and a fair and transparent market to reward unbiased, ethical and human centric AI.

The Progressive Alliance of Socialists and Democrats Group (S&D)

New technologies such as Artificial Intelligence should be developed and used in the European Union in full respect of Union law, human rights, dignity, autonomy and self-determination of every individual, to ensure equal treatment and non-discrimination for all. Possible bias in AI-systems can cause serious harm to individuals and they can perpetuate inequalities in our societies. It is essential to safeguard diversity and inclusiveness in software, algorithms and data used and produced by the AI. We need to promote and invest in solutions to tackle AI bias, through de-biasing of datasets and ensuring data is representative, in order to guarantee the protection of those groups that are vulnerable or might be most affected by discrimination. Multidisciplinary expert teams must be involved in the building of AI-systems. AI technologies should become our ally to eradicate any type of bias or discrimination, to ensure gender equality and the protection of minorities. Responsibility and governance are the key components before any AI-powered product reaches the European single market. We need clear rules and legal standards ensuring non-discrimination and non-biased AI, to enhance safety and promote the citizens' trust.

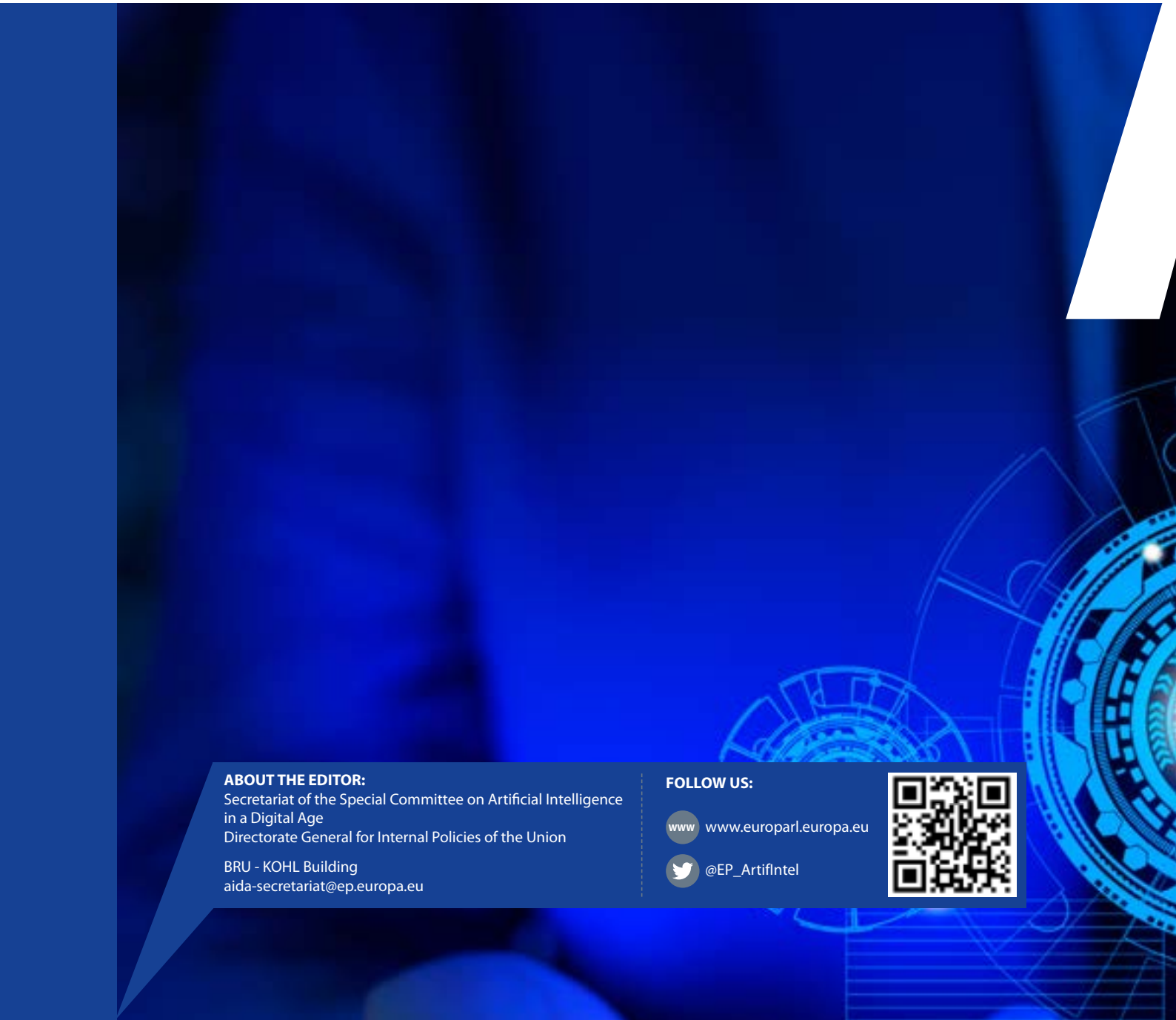
The Renew Europe Group

Renew Europe strongly believes in an approach to AI that is based on the fundamental rights, freedoms and values enshrined in the EU Treaties, the EU Charter and international human rights law. The complexity of algorithms and large data sets may include biases, consequently leading to biased outputs and to discrimination based on social, economic, ethnic, racial, sexual, gender, or disability status or other factors. In this regard, we shall ensure that legislation guarantees everyone's rights, in particular those of the most vulnerable. Any legislation should at least guarantee that structural biases present in our society are not being repeated or even amplified through low-quality datasets. We should consider mandatory human rights diligence rules at early development stage of those AI systems that potentially pose a risk to fundamental rights. Moreover, we need to support development and application of technical means that can contribute to a minimization of such biases and discrimination all along AI deployment.

Greens/European Free Alliance

There is no doubt that practices of AI systems have led to unfair treatment of people in the past, especially those who are already subject to discrimination. This is why it is crucial to highlight the issues resulting from biased technology. However, complex problems, often resulting from inherent societal issues, cannot be solved by simple solutions: while complete and representative data-sets are crucial to ensure that AI systems operate properly, it is essential to prevent the occurrence of bias at every step of the way - from its development to the final results.

In the European Union we have the opportunity to pave the way for fair and human-centred technologies. Fundamental-rights-based assessments can help mitigate some of the risks from AI systems. However, shifting the burden to companies alone will not ensure effective prevention of biased AI and denies the structural underlying issues that are often the root cause. Instead, by empowering citizens in the early stages of AI-development, such as through targeted public funding, we can ensure that European values lie at the core of development of new AI systems and that every citizen can contribute to innovation.



ABOUT THE EDITOR:
Secretariat of the Special Committee on Artificial Intelligence in a Digital Age
Directorate General for Internal Policies of the Union

BRU - KOHL Building
aida-secretariat@ep.europa.eu

FOLLOW US:

www.europarl.europa.eu

@EP_ArtifIntel