SPECIAL COMMITTEE ON ARTIFICIAL INTELLIGENCE IN A DIGITAL AGE

* * *

Public Hearing on

"AI &Bias"

* * *

PANEL I - The impact of bias on the development of trustworthy AI

Michael O'Flaherty, *Director, EU Agency for Fundamental Rights*

Timnit Gebru, *Independent Scholar*

Victoria Espinel, *President and CEO, BSA - The Software Alliance*

Daniel Leufer, *Europe Policy Analyst, Access Now*

* * *

PANEL II - Algorithmic accountability and data governance: How to reduce bias in AI systems

Nathalie Nevejans, *AI & Robotics Law and Ethics Expert*

Dana Rao, *Executive Vice-President, General Counsel and Corporate Secretary, Adobe*

Ricardo Baeza-Yates, *Professor and Director of Research, Institute for Experiential AI, Northeastern University, Silicon Valley*

* * *

BRUSSELS

TUESDAY 30 NOVEMBER 2021

1-002-0000

IN THE CHAIR: DRAGOŞ TUDORACHE
Chair of the Special Committee on Artificial Intelligence in a Digital Age

(The hearing opened at 16.50)

Opening remarks

1-003-0000
Dragoş Tudorache (Renew), *Chair AIDA***.** – Good afternoon, dear Members. Good afternoon, good morning or good evening to our online audience, which I understand is quite broad today, and I'm not surprised, given the topic of our hearing this afternoon. We have the last hearing of the AIDA Committee in 2021, and I think we have a topic for today's hearing that's befitting for the work that we've done throughout the year. I don't think there has been a single conversation that we've had in the course of these months on any given topic that we have looked into – whether it was competitiveness or whether it was democracy, or whether it was health, or whether it was defence, where the element of bias, discussion of bias in algorithms, in the way designers might have cognitive prejudices in the way they do algorithms, in the way datasets can be skewed, either historically or because they are built for purpose – where the issue of bias has not been at the centre of our debate. And therefore we have decided to have this dedicated hearing on AI and bias, and for that we have two wonderful panels: panels of experts, people who have been working and who are working on AI and who have a deep knowledge of how bias may impact the work that is being done on AI and the impact and the consequences and the practical implications of AI in all walks of life.

In terms of the order and the way we're going to organise ourselves, there's no surprise. We will be doing the usual system of five minutes for the panellists, for the speakers, and I will kindly ask them to stick to the five minutes, and then for the Q&A with our Members, we will do the ping pong, two minutes for the question, two minutes for the reply, and I will remind again the Members to be specific about who they address the question to so that we can organise the interaction properly.

We will start with the first panel. I will announce the panellists with the observation that the first panellist has some technical difficulties, and therefore we will start with the second.

So for the first panel, we have Mr Michael O'Flaherty, Director of the EU Agency for Fundamental Rights. As soon as colleagues tell me that his technical problems are solved, we will go to him.

The second, and therefore now becoming the first speaker in the first panel, Ms Timnit Gebru, AI ethicist, and I'm sure you all know her history and therefore we are all very eager to hear her take on AI and bias.

Ms Victoria Espinel, President and CEO, BSA | The Software Alliance, and Mr Daniel Leufer, Europe Policy Analyst at Access Now, a global human rights organisation.

For the second panel, on Algorithmic Accountability and Data Governance: how to reduce bias in AI systems, we will hear from Ms Nathalie Nevejans, Director, Responsible AI France (*Chaire IA Responsable*). Mr Dana Rao, General Counsel and Corporate Secretary at Adobe

and Mr Ricardo Baeza Yates, Research Professor, Institute for Experiential AI, Northeastern University, Silicon Valley.

We will take the first panel, then the Q&A for the first panel, then the second panel and the Q&A for the second panel. So without further ado, I will now give the floor to Ms Timnit Gebru. You have the floor for five minutes.

Panel I

The impact of bias on the development of trustworthy AI

1-004-0000

Timnit Gebru, *Independent Scholar***.** – Thank you for having me. I guess I won't really dwell on what happened to me at Google last year, almost to this day: I was fired very publicly and very disrespectfully. But thanks to a long history of tech worker organising and whistleblowing, they weren't able to smear my work and my reputation as they had wanted to. And so the short gist of my remarks today will be that if we want to reduce the harms of AI, we need to increase worker power and curb the power that these multinational organisations have. That's the only way we can have checks and balances.

So sometimes when I say this, I know that people might get disappointed because they might expect me to talk more about regulation that is more specific to the technology. But I've worked on all sorts of aspects of AI, and I've really come to this conclusion.

Since my firing last year, there's been a lot of developments in whistleblowing and worker organising, and the most famous one, I guess, is Frances Hogan, talking about how Facebook prioritises growth over all else, even when the consequences are deadly. And I saw this happening first-hand. I am born and raised in Ethiopia, and on 3 November 2020, a war broke out, and the effects of unchecked misinformation, hate speech, alternative facts on social media have been absolutely devastating.

For example, I and many others had reported a clear genocidal call in a park on 30 October on Facebook, and the company responded, saying that this didn't violate their policies, and only until there was so much uproar did they, after so much uproar, did they pull it down – and other platforms like YouTube have not even received this scrutiny. Platforms like TikTok, Telegram and Clubhouse were not even discussing it, and Twitter has these issues too to a lesser extent.

So when I was alarmed by these issues and wrote a paper outlining the harms posed by AI models, for example, large language models trained on using data from these platforms, I was fired by Google. So again, when asked what regulations need to be in place, I say the number one thing that would safeguard us from unsafe uses of AI is curbing the power of the companies who develop it and increasing the power of those who speak up against not only the harms of AI, but the tech companies' practices.

So examples of this: the 'silence no more' act has now been passed in California. That makes it illegal to silence workers from speaking out about racism, harassment and other forms of abuse. I think this needs to be universal.

I think that companies need to have much stronger punishments for violating any of these laws, like the huge aggressive union-busting activities that we see by Amazon. When workers have

power, we have checks and balances, where a few billionaires do not do things at their whim and affect the entire world. That's the situation we're faced with today.

So outside of Google, I'm currently working on an independent AI research institute that I hope will have very different incentives from tech companies or the elite academic institutions that feed them. So I'm going to unveil this institute in two days. But working on this institute, I've noticed that the same tech billionaires who run tech companies also advise the government, also have foundations that set the agenda of AI research outside of these large tech companies. So now not only are they controlling the world through unchecked power with these large multinational corporations that are not regulated, they are also controlling all other aspects of AI research as well.

This is not checks and balances, right – so what is the way forward?

In my opinion, to truly have checks and balances, we can't have the same people setting the agenda of big tech research, government and the non-profit world. We need alternatives. We need much more investment by governments around the world in members of communities who attempt to build technology that benefits them rather than the agenda that is currently – I feel that what we have is either the military or big tech.

So I think, contrary to the Cold War rhetoric of the arms race, this is really what stifles innovation, right? Because some people are out there building harmful technology and other people are constantly behind because they're trying to stop this harmful technology from being built. And we cannot implement our vision of the future, because we're constantly trying to stop the cycle. And then once we get to some bad technology like face recognition, then they're on to the next one.

So again, in closing, all I'll say is that only when we change the incentive structure can we see technology that prioritises the well-being of citizens rather than a race to figure out how to kill more people more efficiently or make the most amount of money for a handful of corporations around the world.

1-005-0000

Victoria Espinel, *President and CEO, BSA - The Software Alliance***.** – Good afternoon, Chair Tudorache and members of the Committee. My name is Victoria Espinel. I am the CEO of BSA, The Software Alliance. BSA is the leading advocate for the global enterprise software industry. Enterprise software services, including artificial intelligence, are accelerating digital transformation in every sector of society and the economy, and BSA members are on the leading edge of that, providing European businesses and governments with trusted tools that they need to leverage the benefits of AI.

I commend the Committee for convening today's hearing, and I thank you for the opportunity to discuss the critical impact that bias can have on the development of trust for the AI.

While conversations around artificial intelligence often focus on robots, self-driving vehicles or social media, AI has many applications beyond those. European companies of all sizes benefit daily from AI-powered tools that help improve manufacturing, reduce negative environmental impacts, streamline logistics and detect cyberattacks. One of our members, Autodesk, is using digital tools with AI features to help reconstruct the Notre Dame cathedral in Paris. Autodesk's technology is creating a 3D model of the cathedral using millions of data points. This is helping to keep the reconstruction as close as possible to the original and ensure the long-term preservation of this beautiful monument.

While the adoption of AI can be a force for good, developing and deploying trustworthy AI means acknowledging the risks and working to mitigate them. So my remarks today will focus on three topics.

First, outlining BSA's framework to build trust in AI. Second, offering thoughts on how the European Parliament can improve the AI Act's approach to bias. And third, identifying opportunities for international cooperation.

For BSA members, earning trust and confidence in the software they develop is crucial, so confronting the risk of bias and AI is a priority. Two years ago, we set out to develop concrete steps companies can take to guard against bias. The result is BSA's Confronting Bias: Framework to Build Trust in AI, which outlines a first-of-its-kind risk management approach for increasing transparency and confronting the issue of bias. It is built on three key elements: impact assessments, risk mitigation practices and organisational accountability.

The BSA framework includes more than 50 actionable steps for performing impact assessments that identify risks of bias and corresponding best practices for mitigating those risks. Among the unique features of the BSA framework is that it recognises that these steps need to be followed at all stages of the AI lifecycle: design, development and deployment. Also, different businesses will have different roles throughout the lifecycle. This means that risk management responsibilities need to be tailored to a company's role. Who is developing the algorithm? Who is collecting the data? Who is training the model? And ultimately, who is deploying the system?

What does that all mean in practice? First, when designing an AI system, companies should clearly define the intended use and what the system is optimised to predict. They must identify who may be impacted, and if the risk of bias is present, they must document efforts to mitigate that risk.

Second, at the development stage, companies should document choices made in selecting features for the model and document how the model was tested. And third, at the deployment phase, companies should document the process for monitoring the data and the model and maintain the feedback mechanism to enable consumers to report concerns. And to be clear: at every one of these stages: it is very important for companies to have a team that brings diverse perspectives and backgrounds which could help anticipate the needs and concerns of impacted communities.

The European Commission has addressed many of these issues in the EU AI Act, and we applaud them for focusing on such an important aspect of AI development. The AI Act risk-based approach will help promote responsible practices. BSA looks forward to working with the European Parliament as we examine opportunities to enhance the legislation.

In our view, several important aspects of the AI Act could benefit from a close examination and could be improved in ways that provide greater clarity and greater protection.

One key area of focus is ensuring responsibilities are allocated to the entity that can best mitigate potential harms. Managing the risk of bias is a collective responsibility that involves multiple stakeholders. In many instances the risk of bias may emerge at the intersection of design decisions made by the developer and downstream decisions made by the organisations that use AI. It is therefore important to ensure that compliance responsibilities are assigned in a manner that reflects the different roles of AI developers and AI deployers.

Finally, rules that mitigate the risk of bias will be most effective if they can be applied globally. BSA strongly encourages international cooperation on AI with like-minded partners, both in regulatory approaches and, more specifically, with regard to mitigating the risk of bias. The US

Administration recently called for an AI Bill of Rights, which may fit well with the EU imperative to protect fundamental rights. An approach that is consistent and interoperable among like-minded democracies, built upon the common ground of common values, will benefit consumers and will ensure that all developers and deployers of AI adopt best practices for assessing and mitigating risk.

We appreciate the Committee's focus on these issues. We hope the BSA's framework on confronting bias will be a significant contribution to this and ongoing discussions in the European Parliament. Thank you again for the opportunity to participate in today's hearing, and I look forward to your questions.

1-006-0000
Michael O'Flaherty, *Director, EU Agency for Fundamental Rights***.** – Chair, Honourable Members, we're grateful to be with you, and my apologies for the technical glitches at the beginning.

Honourable Members, bias, prejudice and preference are as old as humanity, and they are integral to – often the cause of – some of the great evils that we experience: racism, anti-Semitism, just to take two. Frankly, it was inevitable that bias would migrate to human-made technology and decision-making, and we see that it is throwing up problems right across the human rights spectrum. We've seen, for example, the impact for access to social welfare payments. We've seen the problems when biometric ID systems are deployed by security forces.

Today I would like to narrow my opening remarks specifically to the relationship of bias and discriminatory acts – or in other words, automated decision-making that triggers unlawful distinctions between people on the basis of the so-called protected grounds. As a matter of law here in the EU, those protected grounds include sex, race, colour, ethnic or social origin, language, religion, political opinion, property, disability, age and sexual orientation. And I recall again, that is not a closed list.

Now, we're familiar with the various manifestations of such discrimination: algorithms that prefer men to women, that prefer one skin colour over another. We can also see intersectionality in play, where different discriminations interact. For example, an AI bias that triggers discrimination against socially-disadvantaged people and that also then favours the young or the old. This phenomenon deserves serious attention in and of itself. It's a critical human rights issue, but also, taking the topic of today's discussion, it's also all about trust. How can we trust AI that does such harm?

So briefly then from me, some specific action points. Just six, but in headline fashion.

In the first place, apply what we already know in tackling such curses as racism. Work with impacted people, never for them. Foster diverse workplaces in industry, in oversight and in the policy spaces so that the people impacted by bias are there: cognizant, aware and alert.

Third, we need to acknowledge the extent of the problem we're facing and of the remarkably low levels of awareness of this extent. We see this in the general public. We know this through our work – our fundamental rights agency research – but we also see it in the tech sector, where we see, for example, a number of misconceptions.

Just two examples. The misconception that AI with no protected characteristics data cannot discriminate. This is simply not true. We can have data in there closely related to a protected characteristic that serves as a proxy that triggers the bias, such as, for example, how shopping behaviour can be a proxy for gender.

Another misconception is that bias can be easily fixed by de-bias datasets. We're not impressed on the basis of the scientific findings on this claim. Third, we need to take account of and engage the legal complexity here. Discriminatory acts can only be assessed in context. Bias can only be perceived to be discriminatory in the specific context, and we have to acknowledge that in our responses and of course, within the EU space, we have to acknowledge where EU competency begins and ends.

But fourth, in terms of fixing the problems, we can achieve a lot with a focus on data quality issues. We need to avoid training data that promotes discriminatory behaviour. Avoid under-representation of groups in training data. Avoid humans exhibiting bias in labelling training data – babies and black babies, to take one famous example of labelling. And of course, we also need to watch how feedback loops may promote bias. This is something we are actively researching at the moment in the agency.

And then fifth, it's imperative that we create the machinery, the space for an assessment of the quality of AI. We need, in other words, to ensure the putting in place of ex-ante human rights and fundamental rights assessments, at least with regard to high-risk AI applications.

Now, while insisting on the need for such assessments, we do acknowledge the difficulties in putting an effective system in place. We recognise that it will require transparency if it's to be trusted, but that transparency throws up its own issues, albeit we insist on it. And secondly, we recognise the challenge to develop the necessary training data to do the assessments that take account of information on human protected characteristics, at least when the private sector tries to do this. So we acknowledge the challenge, and I appreciate that it's anticipated in the draft AI regulation.

And then finally – sixth and finally – we would urge that we all stay heavily invested in the need for research in this area. There's a lot we don't know. We will continue to need to learn, and that learning must be an interaction of technology, ethical specialists and experts in the binding human and fundamental rights obligations.

I'll leave it there for now, and I would just like to assure the honourable Members that you can continue to count on the fundamental rights agency as a partner as we face these great challenges.

1-007-0000

Daniel Leufer, *Europe Policy Analyst, Access Now.* – Hello, and let me just start by saying that it's really an honour to be asked to speak here today and to be in the company of such distinguished panellists, especially Dr Gebru, of whose work I'm a massive fan.

We're here today to talk about the impact of bias on the development of trustworthy AI. Access Now has been working on artificial intelligence for a long time now. We were one of the first civil-society organisations pushing for a human rights framework to be applied to AI when all the talk was about AI ethics, and along with Amnesty International and some other partners, we drafted the Toronto Declaration in 2018.
I'd like to talk about the topic by reframing it a little bit. So rather than talk about the impact of bias, I'd like to talk about the impact of framing the problems caused by AI systems only in terms of bias and how that affects the development of trustworthy AI.

I think the first impact of this framing of everything in terms of bias is to distract from structural solutions to problems. I'll take an example (and there are many examples). One of them is AI proctoring software. This is software that's used to determine if students are cheating on exams, for example, and we saw an explosion of software like this during the COVID-19 pandemic.

It's been shown that these systems are biased against students with darker skin, that they actually fail to even detect the faces – the most basic step in the facial recognition system – even detect the faces of students with darker skin, so that black students were forced to shine incredibly bright lights on their faces throughout exams.

Now obviously, if we're using a system like this, we do not want it to be biased against certain groups; we don't want to have situations like this. But is an unbiased version of this type of surveillance software something that we want in our technological future? I think the answer is no, that we don't want to be moving towards ever greater surveillance, and I think we need to be aware that often, bringing AI into a situation means bringing surveillance. It means bringing data collection. So we need to be careful about that.

The second impact of this bias framing is to shift governance and power to tech companies, further cementing their power. If we take, for example, if universities become reliant on AI proctoring software, it's likely that it will be procured from third parties. That means that decisions about what constitutes suspicious behaviour, what's bias, what's discrimination, will be shifted into this technical domain and can be taken out of the hands of our public institutions. So we need to be really, really careful to avoid that. And this is part of a general shift in seeing who's an expert on a topic rather than a human rights lawyer – does it become a machine learning engineer who's the new expert on bias? So we need to push back on that.

In the time that remains, I'd like to make some comments on the EU's AI Act, to discuss some proposals from civil society that push back against this tendency I've just outlined.

Under the current proposal, most of the obligations fall on what are called providers of AI systems that are, you know, companies who are putting systems on the market and developing them. And they have to do some quality checks as part of a conformity assessment, and in there, there are some things about ensuring that there is no problematic bias, that data sets are of high quality.

Now this is a valuable step and in itself is good, but it will not be sufficient to protect people from the range of harms and oppressive impacts that can come from AI systems. So there's two prongs that we're approaching this problem from.

On the one hand, we need to ensure that the list of prohibited practices in Article 5 is expanded and strengthened and can be updated in the future so that the most dangerous applications of AI – things like emotion recognition, predictive policing and, to come back to what Mr O'Flaherty mentioned, systems that we've seen that claim to be able to detect someone's sexual orientation from their facial features or their political orientation from biometric data – these systems need to be prohibited outright so that people can trust that there isn't this free-for-all where any kind of crazy rights-violating system can be placed on the market.

Then the other thing that we need to look at is, for the high risk systems that are going to be deployed, how can we ensure that trustworthiness can be guaranteed?

We need more obligations on the entities deploying and using AI systems, so users under the AI Act. We want them to do fundamental rights impact assessments, which involve assessing the impact of their systems and deployments on fundamental rights and to identify people who are likely to be affected by those systems.

This is part of a big problem with the AI Act: that it doesn't actually confer rights on the affected people. So civil society and, today, Access Now, along with the 114 other civil society organisations, released a statement where we outlined the issues I've just raised and more, about

how we can make the AI Act work for fundamental rights. So Access Now, European Digital Rights and the other signatories remain open and willing to talk about how we can ensure that this AI Act really works to protect fundamental rights, and I'll drop a link to that statement in the chat.

1-008-0000
Eva Maydell (PPE). – Thank you very much for having me, and I hope you could hear me well, and I'm glad to hear those interesting contributions. I got triggered by a couple of things that were mentioned by Mr O'Flaherty, and I would like to know in how far you see basically the issues surrounding bias in AI as radically different to cases of bias in tech in the past?

I'm happy to see whoever wants to contribute – Ms Gebru or any one of the other panellists – because when I look back to the 1950s, new tools such as refrigerators or washing machines freed many women from hard manual labour. But many of these household tools in their marketing also reinforced this sort of traditional picture of women. You surely know that the advertising that was used then sounds silly today.

So either these were mistakes caused by sloppiness or negligence, or they were intentional acts of misogynism. So if that is the case in AI too, then we have the tools, because you can fight sloppiness through quality control and standards, and you fight hatred or prejudice through laws, through societal debate and through the power of a transparent market where customers can choose. But your thoughts – again, how far you see this issue surrounding AI bias as radically different to cases of bias in tech in the past – will be very interesting for me to hear.

1-009-0000
Chair. – Thank you Ms Maydell: any preferred panellist for the answer to your question, or I will choose myself?

1-010-0000
Eva Maydell (PPE). – Perhaps Ms Gebru, but also it would be interesting to hear from Victoria Espinel as well.

1-011-0000
Timnit Gebru, *Independent Scholar***.** – I would say that there are many ways in which it's similar and some ways in which it's different. For instance, my colleagues have written about what they call runway feedback loops, where the bias – quote unquote 'bias' (and I agree that this bias framing can be problematic) – that is built into technology, into machine learning in particular, can widen already existing bias because of the what we call runway feedback loops. So it can make existing issues even worse and it can entrench and centralise power much more, and it can increase that existing discrimination and historical issues that we are experiencing. Even in my paper, for example the last paper I wrote, we talk about how, even when we go past these social issues, these models that we train on historical data can take us back, because they don't track the way in which society has been overcoming or trying to overcome these issues. So I can talk about this for an hour, but I'm just going to leave it there for now.

1-012-0000
Victoria Espinel, *President and CEO, BSA - The Software Alliance***.** – So just very briefly, I would say the reasons for bias in the past and today are manifold: sometimes accidental, sometimes it's negligence, sometimes it's intentional. Regardless, I think, of what the origin story of that is, I think your point, Ms Maydell, that there's a range of tools is exactly the right one, and I think what we all need to be doing collectively is working as hard as possible to bring the full power of that range of tools to bear.

I will say one thing that I think is different today than compared to the 1950s, say, is that there is much more focus on this now. And so that is not, you know, that does not mean that the job is done, it means it's only beginning. But for example, the fact that we're having this hearing,

the fact that there is legislation, I would bet that there was not legislation coming out of the European Parliament addressing issues of bias and advertising in the 1950s.

So I think the fact that there's more awareness is very positive. Now, obviously, the challenge is to take that awareness and what can we do in a very practical way to try, to the extent possible, to mitigate the risk of bias. And I would echo something that Mr Leufer said, not just on developers – including developers – but also on designers and, really importantly, on deployers, on users of AI in the AI Act parlance. So I think, looking at the full range of tools that we have, whether they're technical, legal awareness and others, using as wide a spectrum as they can is critical.

1-013-0000
Ibán García Del Blanco (S&D). – Further to the last speech, I agree that we should choose a mixed system for assigning responsibility, and thus to try to ascertain what intervention is required for such a complex product in each of the steps needed to place it on the market.

But it is also true, however – and I am addressing Ms Espinel here – that it seems very complicated to assign this by means of a legal regulation, due to the inherent complexity and the differentiation that we would have to make between each of the aspects. I do not know if she shares my opinion as to the difficulty.

And there is another issue I would like to raise with Mr Leufer, which has to do with the fact that, when we talk about bias and discrimination, we are of course talking about human rights. I don't know whether or not he shares the view that the Commission's proposal for an artificial intelligence act pays sufficient heed to this issue when differentiating between risk systems.

In other words, the location of each of the applications will have to do with the potential risk. But, of course, when we talk about human rights and discrimination and bias, we are ultimately talking about fundamental rights. In other words, how is it possible for us to keep some applications out of the scope of this prior hearing regardless of their final use, when they could potentially commit discrimination?

I would like to ask Mr Leufer if he thinks this proposal sufficiently protects the application of artificial intelligence or technologies using artificial intelligence on the European market or not.

1-014-0000
Daniel Leufer, *Europe Policy Analyst, Access Now***.** – If I understand the question, it was basically whether the risk-based approach as it exists in the AI Act does enough to protect fundamental rights. And I think from the beginning we were critical of the idea of taking a risk-based approach, and we pointed out some potential flaws with that. One of them would be that it would not include a category of unacceptable risk, which thankfully the AI Act does, although we have, you know, criticisms about exactly what's currently listed in the prohibited practices.

I think the real point to get at is: is it possible in advance to know what type of risk different applications pose to fundamental rights? And we don't think it is. In the AI Act as it stands, there's a possibility to update the list of high risk systems, so in Annex 3. But there's no corresponding possibility to update the list of prohibitions or the list of what's called limited risk systems under Article 52.

So one thing that we've asked for is for an update mechanism, and we need to have a discussion about exactly what that would look like, to be included to allow for new prohibitions to be added, to allow for systems to be reclassified to a different risk level if evidence emerges that shows that actually they pose a different threat – a more serious threat than was first perceived. And I think, if the risk-based approach that we have isn't dynamic enough to allow for those reclassifications, then it will become outdated, will not be able to keep up with actual threats.

I think another key point there is to have a proper flagging mechanism for people who are affected by systems. Under Article 64 there is a possibility for fundamental rights enforcement bodies to investigate systems, but there's nothing there that would allow affected people to actually flag that they think they've been affected, that their rights have been violated or undermined by a system.

So again, with the document that I put in the chat, we've outlined some proposals for how to make that risk-based approach work better for protecting fundamental rights.

1-015-0000
Karen Melchior (Renew). – Thank you to the panellists for being here. It's been very enlightening to hear your thoughts, and thank you especially to Timnit Gebru for her courage in taking a stand with Google as a former employer. I was very inspired by both Daniel Leufer's and Timnit Gebru's points on how to improve the development of AI and the tech companies, also the rights-based approach in Access Now and the Fundamental Rights Agency.

However, looking at the legislation, how can we improve on the AI Act to take into account a more rights-based approach and also to make sure that we look at all of the steps of AI development, as was outlined by Victoria Espinel?

One of the concerns that I've heard is that there is not enough realism in the way that the AI Act has been drafted. For example, we cannot have perfect data sets in order to prevent bias. So I would like to hear Ms Espinel for her recommendations in improving the AI Act.

1-016-0000
Victoria Espinel, *President and CEO, BSA - The Software Alliance***.** – Thank you very much for the question. I think in terms of specific suggestions on the AI Act, here are a couple that I would offer.

I think in terms of allocation of responsibilities, the AI Act, as Mr Leufer said, does have responsibilities for the developers of AI. I think it's important that we acknowledge that responsibilities need to be allocated throughout the entire lifecycle of AI. And what I mean by that is there need to be responsibilities allocated on developers, on designers and also on deployers – on those that are using the AI system. I think having the entire ecosystem of those that are creating but also using AI is very important. And so having the AI Act also apply to all levels, all different roles and responsibilities in creating and using AI, I think is very important.

The second thing that I would say is, as I mentioned, we spent two years of very intense work on the BSA framework, very interested in feedback and very conscious of the fact that we will be working on an ongoing basis to update it and improve it. And with respect to the AI Act, I would say similar. It's very commendable that the European Commission has focused on this issue early and very seriously, but it is also true that situations will change. And so again, echoing what some of my fellow panellists have said, I think it's important to have feedback mechanisms, and I think it's important to be drafting now, to the best that we can, but conscious of the fact that things will change and this legislation will need, I imagine, to be updated or re-reviewed to make sure that it is state of the art, that it is addressing the problem as well as possibly can be addressed, both now and into the future.

1-017-0000
Alexandra Geese (Verts/ALE). – First of all, Dr Gebru, thank you for joining us today, it's really a great honour. Since you said you wouldn't want to comment on regulation that much, especially European one, a question on something you said.

You said public investment and civil society organisations would be important. What kind of instruments do you envisage there? I'm quite happy because the new German Government's

coalition agreement has exactly that: funding for civic tech, for diversity developed by civil society. So is that the direction where you would see that going, or do you have any other ideas?

And another question on workers' power. That's absolutely correct, and I think in Europe we might have some better instrument to protect workers in those cases. But the problem with big tech, as you very well know, is that the people working there are not exactly representative of a diverse society. So what can we do to promote that?

And then a quick question to Daniel Leufer. How do you think we can involve affected groups also in the development of artificial intelligence, or where could we anchor that in the AI Act?

1-018-0000

Timnit Gebru, *Independent Scholar***.** – Thank you so much for this question. In terms of investment into civic society, what you're mentioning is exactly one of the things that I'm talking about. Currently the only people who have the resources and the time and the backing to develop technology are basically either – the source of funding is either – the military, so there's this arms race, or large tech companies. And so if the goal is to either do one or the other, then when we percolate backwards, we can't expect a different outcome than what we have right now.

So if we want a different outcome, we need to invest in specific goals for the development of technology that are not tied to these two institutions. Because what is currently happening when we talk about impacted communities, impacted communities cannot work on their version of the future. They're constantly behind. The burden right now to prevent harmful AI is on them.

When you look at social media companies, etc., the burden of the work is on impacted communities to show that there are issues. And then by the time they're exhausted showing those issues, then the companies or other people have moved on to the next thing. And so we're constantly running behind. So that's why I really believe that this proactive investment is really, really important.

And I think so, for example, you were talking about how to have more diversity. I think that proactive investment is one of them, but reactively having higher punishment for the kinds of things that, for example, Google would do to me (they do it to many others) is very important, because if we don't have that, then all of these frameworks we're talking about, about who is responsible at which part in the AI development process, won't be implemented, because the people who would implement them are pushed out like I was.

I can say a lot more, but I'll keep it short. But those are my two comments on this question.

1-019-0000

Daniel Leufer, *Europe Policy Analyst, Access Now***.** – Thank you for the excellent and extremely difficult question. And to be honest, I think Dr Gebru has answered parts of it already, and I can't add much to what she said.

Within the context of the AI Act, I do think it's difficult and I do think we're actually in a way trapped in what Dr Gebru just spoke about, in that we're looking at how to deal with the harms that are going to pop up. Unfortunately the Act is not ideal and it's put us in a position of trying to, like a game of whack-a-mole, figure out how we can best identify harms and deal with them.

So I do think there's probably not that much within the AI Act that we can do there. But one thing that we can do is to make sure that providers and users under the Act have to identify affected groups – that the Act actually gives rights to affected groups – so that we at least move away from the terrible situation we're in now, where there's no consultation, where there's no transparency. We don't know what's being deployed, and unfortunately, what we need to do

with the Act is get out of that situation and then we need to look at other mechanisms, I think, like public investment in publicly-owned and -developed technologies that are really participative and make sure that we push back against this tendency to have public institutions and services outsourced to private companies.

1-020-0000
Alessandra Basso (ID). – Mr President, dear colleagues, thank you to those who have joined us for their valuable input.

My first question is for Mr O'Flaherty: when talking about the digital world and artificial intelligence, one of the biggest concerns is undoubtedly disinformation, especially when political in nature and spread by external actors. How can we tackle this problem, and large-scale organised attacks using AI in particular?

I also have a second question, which Ms Espinel may also be able to answer, as she spoke about anticipating the needs of the impacted communities. What can we, as legislators, do to protect the most vulnerable, such as persons with disabilities, from possible prejudice arising from the misinterpretation or misuse of available data?

1-021-0000
Michael O'Flaherty, *Director, EU Agency for Fundamental Rights***.** – Yes, tackling disinformation and misinformation is a huge concern. There are a number of technical supports that are being tried out by social media platforms right now; others are better placed than me to comment on their quality. But we're investing heavily in the business of fighting lies with fact, of making sure that we have a strong evidence base to refute the nonsense that does so much harm. And we have to, by the way, fight a long game in terms of the response. The truth can eventually win out if we stay sustained and smart and clever in our messaging.

We also have to use criminal law more vigorously. When misinformation and disinformation crosses the line of criminal culpability, it needs to be investigated and prosecuted – something that's not being done often enough right now.

And if you'll allow me, Chair, just one comment on regulation: the question the other panellists have the opportunity to reply to. I associate myself and my agency with much of what we've heard. I just want to insist on two further dimensions.

One is transparency. It is critical that the data be made transparent, that the assessments and the content of the assessments be disclosed, that the training data is disclosed and right across all of the dimensions. That's the only way we can have confidence and trust in AI – and we, by the way, will also have to strengthen oversight. As the European Data Protection Board and ourselves have said, the supervision system that's currently envisaged is quite good in terms of its protection of privacy, but we need a broader human rights capacity at that level of independent oversight.

1-022-0000
Victoria Espinel, *President and CEO, BSA - The Software Alliance***.** – So thank you very much for the question. I think in terms of what legislators can do, you are doing at least part of that right now – having hearings like this and drafting legislation that focuses on this issue is the critical role of legislators. I applaud you for having a very active process for getting input and working hard to make the legislation as good as it can be at the moment.

As I have said, we have a few suggestions. I think others have made suggestions for how the AI Act could be improved, and so I would urge all the members of the Committee to look at those suggestions seriously. I do think that having obligations on impact assessments and having obligations that for companies, both users and developers of AI, to be thinking about communities that could be impacted is an important part of this. That is something that

companies should be thoughtful of, regardless. But I think, as you are in the process of drafting legislation, having that be a component of the legislation, I think, is really critical. As I said, it is something that we would hope would be done regardless, but it may not be.

And so I think that your role as legislators is to help draft legislation that is as thoughtful and future-looking as it can be, including creating some important obligations on the developers and users of AI. And then again, I would also urge you to continue focusing on this issue into the future. What I mean by that is the legislation will, like all legislation, be tested by time. This is an issue that will continue to evolve. Hopefully, efforts like yours will help mitigate and reduce the risk of bias, but I think there will be more that can be done in the future. So I commend you for what you're doing now and ask you to continue to have this be an issue that you are focused on.

1-023-0000
Kosma Złotowski (ECR). –. Good afternoon. I also have a question for Mr O'Flaherty. When we talk about discrimination in AI systems we are mainly talking about decisions that are the result of inputting data that might not reflect reality. Then we can talk about it being the result of bias on the part of the system's authors or the person who compiled the data set. Discrimination is thus a result of human error or deliberate action.

But what if an AI-based decision-making system has available to it the most objective possible data, which has been repeatedly checked and structured, yet the end-effect favours or discriminates against a particular group? Can we say that any result produced by an AI system that corresponds to our idea of discrimination really is discrimination? In so doing are we not manipulating the technology and somehow forcing it to make decisions that are not objective and based on data, but which are in line with our social or political views?

1-024-0000
Michael O'Flaherty, *Director, EU Agency for Fundamental Rights***.** – First, let me agree with Mr Złotowski that you can get very bad outcomes when you put bad data in. We've seen this in migration. We've seen, for instance, where a person was marked up as an adult instead of a child, with terrible consequences in the interoperability of the large-scale EU databases. So it's a serious issue.

But in terms of making a determination of when the outcome is discriminatory or not, it's not a subjective opinion. It's not about values, it's a matter of law. And for us, it's a clear matter. It's only a matter of discrimination when it violates European law relating to discrimination – that means the protected grounds. I didn't invent the protected grounds I listed earlier, I took them from the EU Treaties and from the Charter of Fundamental Rights.

And so at least for us, it's a clear matter. And it's important, in fact, that we can always trace our claims of discrimination, our condemnation of discrimination and our action with regard thereto to revision in the legal commitments.

1-025-0000
Pernando Barrena Arza (The Left). – Good evening everybody, thank you Chair. Governments in the Member States are increasingly adopting AI-powered algorithms to rank their citizens and grant social services, either civil rights or *(inaudible)* services, according to that ranking. Austria, for example, has AI-powered algorithms to offer social services, including scoring people based on their employment prospects and prioritising services based on that ranking.

In the Netherlands, the government used algorithms to penalise people, predominantly in lower-income neighbourhoods, based on whether they were likely to have committed benefit fraud.

In this context, I would like to ask Mr O'Flaherty: could the scope of AI high-risk lists be enlarged to those sectors? How do we ensure that the algorithms used by the public services themselves are not biased? What sort of intervention is required in the public sector to ensure that artificial intelligence systems that are used are ethical, unbiased and do not penalise the most vulnerable in society?

1-026-0000
Michael O'Flaherty, *Director, EU Agency for Fundamental Rights***.** – Yes, the key is risk assessment. It's vital that every AI application with potentially serious impact for human wellbeing is subject to an ex ante risk assessment. And indeed, as a speaker said earlier, let's make sure to avoid closed lists of the different categories of risk. We only become aware over time of the levels of risk that are engaged in a particular application, and the issue of deprivation of social welfare is a matter of huge human consequence. And so it's vital that any such application going forward would be subject to a sturdy risk assessment of the type I spoke of earlier.

1-027-0000
Ernő Schaller-Baross (NI). – Dear Ms Espinel, dear Mr O'Flaherty and whom it may concern, online social media platforms use AI algorithms to create digital profiles of users and deliver content that conforms to their preferences. This practice makes people become separated from information that disagrees with their viewpoint, effectively isolating them in their own cultural or ideological bubbles. Machine-learning algorithms and popular subsets of artificial intelligence are especially good at finding patterns and correlations across large datasets.
A recent Cambridge Analytica case shed light on the dangers of the bias and granularity of control that the AI algorithms provide over the content that users see. There are other cases, also in the field of targeted advertising, where the biased operation of social media algorithms has created discrimination and excluded or restricted certain social groups from accessing information. There are those experiences with the bias of social media platform algorithms.
What kind of EU level policymaking and regulatory solutions do you think would be available to avoid artificial intelligence bias across social media platforms?

1-028-0000
Michael O'Flaherty, *Director, EU Agency for Fundamental Rights***.** – Thank you very much. In response to Mr Baross, it's clear that a voluntary code is inadequate to address the serious issues that you mentioned. I can only agree with much of what you said. And so therefore the DSA – the Digital Services Act currently in draft – is critically important. We have to get this right. We have to strengthen oversight in the area of digital services. We have to embrace models with both external strong oversight and self-policing. And we have to develop the machinery, including finalising work on the Digital Services Act with human rights holders intimately involved. And the key to that is making sure that there's always a place at the table for civil society.

I'll leave it there, Chair. I just want to thank you for the invitation. You cannot imagine how important the issue of AI is to the Fundamental Rights Agency. AI is about the future of our society, and if we don't get it right, it will be a society nobody will want to live in.

1-029-0000
Timnit Gebru, *Independent Scholar***.** – My message is that we need to curb the power of these multinational corporations. I've written papers that have made it into some of the draft legislation, and I'm grateful for that. And I agree with all my other panellists. But I see you mentioned social media. I am just so terrified. I don't have all the time to explain the reasons for the lack of accountability and all of the issues that, for example, social media is proliferating.

We speak about AI. They don't even have language technology for many of the languages people are using on social media, so they can't even use hate speech detection or flagging. But there is absolutely no regulation that stops them from operating as they are currently operating.

So my message is: thank you for inviting me. We just really need to curb the power that these multinational corporations have.

1-030-0000
Victoria Espinel, *President and CEO, BSA - The Software Alliance***.** – First, I want to thank you for having me, and more importantly, thank you for focusing on this extremely important issue. I will very briefly leave you with three thoughts.

First is, it is incredibly important that there be a broad range of stakeholders as part of this discussion, both in terms of drafting the legislation but also in terms, as we have said, of creating and deploying AI. So throughout the legislative process, but importantly, throughout the use of AI, a broad range of stakeholders is absolutely essential.

The second thing that I would say is that, looking at the different roles that are incredibly important (this point has been made by a number of panellists, but I just want to emphasise that one more time, because of its importance), looking throughout the whole AI lifecycle and making sure that roles and responsibilities are allocated appropriately throughout is very important.

And the third thing I would encourage is continual focus on this important issue. Thank you so much for having me. It was an honour to be here and an honour to be with my fellow panellists.

1-031-0000
Daniel Leufer, *Europe Policy Analyst, Access Now***.** – Just to finish, I would really echo the call from Mr O'Flaherty about including civil society, and I'll point to two things there.

One is the call for a ban on remote biometric identification in public spaces. We have the Reclaim your Face campaign in the EU and Access Now runs the Ban Biometric Surveillance campaign, which has been signed by over 220 civil society organisations from over 65 different countries, which all agree that this technology – facial recognition, other types of biometric surveillance in publicly accessible spaces – is incompatible with fundamental rights. We need a full prohibition on that.

The other thing I would say is, please, to all policymakers, reach out, engage with us. We have a coalition of 115 civil society organisations outside of the normal digital rights organisations – really a broad coalition – that have settled on a number of demands for the AI Act to make sure that it works for fundamental rights. We're here. We've done our reading. We're willing to engage. So let's do our best to make sure that this Act really works for fundamental rights and really, really actually makes trustworthy AI a thing in the EU.

Panel II

Algorithmic accountability and data governance: How to reduce bias in AI systems

1-032-0000
Nathalie Nevejans, *expert in AI and robotics law and ethics***.** –I'm going to be speaking about artificial intelligence, bias and legal issues. For a long time it was thought that technologies were neutral. Today, with AI, we are realising that technologies are not neutral, and we have many examples of this. We already know bias comes from a variety of sources. We see it in training data, for example, and in the models themselves. We can find bias in results and in AI

decisions. So today we're realising that it's vital to tackle this bias. But how? The problem is that bias in AI stems from the perceptions of both the designer of the system concerned and society as a whole. So in fact it is the mentalities involved that have to change, for example by means of education. If we take a closer look at the legal issues at play here, we can see that the technological solutions we need to detect bias in AI already exist. And we'll need to incorporate those solutions into legislation in order to combat bias. There are two points here.

In the first instance we're going to see technological solutions that address the legal issues involved in countering bias in artificial intelligence. Today, we have a number of technological solutions that have been developed to counter bias, and they can be used by the operators concerned, including providers. We have systems and solutions that have been developed by companies, and solutions that have been developed in the research sector. I have three comments on that point.

First of all, it is very costly for companies to make AI models fairer. We have to bear that in mind. Secondly, therefore, there is a risk that self-regulation with regard to bias will be complex or even impossible. And lastly, without regulation, it is understandable that businesses – as mentioned earlier — will try to avoid wasting time and money on the issue of bias.
My studies of these different technological tools have shown that they vary in scope and effectiveness. For example, there are tools that will apply only to data. Then there are other tools that will work on the model as well, others that will look at the outcome and others still that will look at the whole decision-making process from input data to results. So there is a host of possibilities that vary depending on the sector in which they're used. This means, for example, that it might be difficult for the operators concerned – in other words the future provider who's going to put the AI system on the market – to decide which one to choose.

So I think the EU could have a role to play in helping give the operators concerned a clearer picture, for example by coming up with a certification scheme or something else – although certification might be a good idea – to validate a certain number of tools that meet the required standards.

The second point is the need for the European institutions to come up with legislation on bias. Bias exists in various forms. It has an impact on society, which is clear, as we mentioned earlier. And, as we've just heard, self-regulation is weak. Each of those factors shows that the EU now needs to legislate on this issue. We shouldn't rely on self-regulation, as some people have suggested – not today, but on other occasions.

The EU has already started working on the issue of bias, which is really good. The main example is the 2021 proposal for a regulation that we've just mentioned. There are two interesting examples in the proposal, specifically in Article 10, which provides for an examination in view of possible biases, and Article 15(5), which refers to appropriate mitigation measures.

I have three comments to make on the proposal.

First of all, I think we need to define the concept of AI bias that we want to combat. The reasons for this is that distinguishing between different kinds of bias can be complex. There is 'unfair' bias – we've spoken about discrimination on the grounds of race and gender. There there's 'fair' bias, which can be corrective adjustments made to a system, for example adding points to the score obtained in a certain category – to recruit more women, perhaps. And then there's simple mathematical bias that scientists use to affect the point at which a neural network is activated.

The proposal should indicate how far operators need to go in their research to eliminate, or limit, bias. Article 15(5) refers to mitigation. But how far should that mitigation go?

Last but not least, the text refers to bias only with regard to 'high-risk systems'. As far as I'm concerned – and you've already mentioned this – a risk-centred approach like this might imply that there's no need to look into bias in other cases.

So we can see where the risks are and where the high risks are. But as far as moderate risks are concerned – chatbots for example – the impression is that there's no need to look into bias. So in my view, combating bias is something we in Europe have to seize upon as a principle, regardless of the risk level involved. And this is something you have already mentioned.

I'll conclude in one minute, but very briefly, three conditions need to be satisfied to gain the trust of the European public.
First they need to be sure that AI is taking fair decisions that respect their rights and values. They must also be able to understand how those decisions are taken. And lastly they should always be able to appeal against a decision taken by AI or by a human assisted by AI.

So the issue of bias is an extremely important one, but it's just the tip of the iceberg. We have a lot of hard work to do on issues like transparency and the explainability of AI.

1-033-0000
Dana Rao, *Executive Vice-President, General Counsel and Corporate Secretary, Adobe***.** – I appreciate the opportunity to have to speak to the Committee. At Adobe, we believe in the promise of artificial intelligence. We today will bring you some practical experiences we have with implementing AI and some of the measures we've taken on self-regulation to help inform the discussion as you consider this bill.

At Adobe, AI powers amazing things across our product portfolio: streamlining, image editing capabilities in products like Photoshop, automating digital document workflows so you can do in seconds what used to take days, and helping enterprise customers gain data-driven insights and target their marketing for better results. With its focus on content-based productivity, Adobe's AI-powered products would fall within the non-high risk categories, as outlined by the EU's approach. However, even for companies whose AI is primarily low risk, it's important to continue to strive to make AI better for everyone by combating potentially unfair, discriminatory or harmful results.

Guided by our principles – Adobe's principles for AI – of accountability, responsibility and transparency, Adobe has created a comprehensive engineering structure internally, overseen by our AI Review Board, to ensure that we are working with datasets that represent society as a whole. We're auditing our findings that verify that we have captured the intended diversity and that we're transparent about our process. Over the past year, we've reviewed over more than 50 product features, helping to ensure our AI is developed, implemented and maintained within the framework of our AI ethics principles.

We welcome the European Parliament's efforts to address bias in AI, but we believe efforts to reduce bias in AI should focus on preventing harmful outcomes rather than solely focus on trying to avoid bias entirely. As a technical matter, it is not possible or even interesting to achieve a completely unbiased dataset. After all, the point of an AI is to learn about tendencies that exist in the data and act on them. For example (someone had mentioned targeted advertising before), if you would like to market hats to women, it would make sense to have a dataset with more examples of women than men to understand their preferences in hats. The bias toward women in that dataset is the very point of building the AI system and does not represent anything harmful.

However, if the dataset is for loan administration, the data has to represent fairly all demographics or the AI could create unfair outcomes. Additionally, harmful bias exists relative to the context of how the AI is being used in a particular use case. In other words, AI that is considered harmful in one specific use case may not be harmful in another. A dataset of hat-wearing women is fine for hat marketing purposes but may not have the right diversity to train a different kind of AI algorithm. This is why we believe any regulation on AI development should be process- and transparency-oriented and focus on testing the outcomes of AI systems before deployment and applying safeguards against biased outcomes after deployment.

We do agree with the EU's focus on increased regulatory efforts on high-risk AI, without slowing down, however, the pace of innovation around AI that poses little or no risk. This will allow low-risk AI innovation to proceed to market quickly but continue to provide safeguards in those areas where fundamental human rights are involved. For high-risk AI, we do agree that the incorporation of human oversight as part of the final decision-making process is critical, as we cannot yet rely on AI to make these complex, nuanced decisions on its own without bias or discrimination.

As you continue to develop the EU's approach to AI, we urge policymakers to ensure that the AI Act contains future-proof, efficient and technically-feasible provisions when it comes to addressing bias in AI. It is also important that policymakers continue to assess the relative tension between privacy laws, which are focused on the minimisation of data access, and AI laws, which require more data in order to allow AI to produce the fair and accurate results we all want. For any given use case, the larger the dataset, the more fair the AI will be. So policymakers should continue to prioritise ways to enable innovators to obtain clean, privacy-friendly data so we can build fair AI.

Finally, we urge policymakers to leverage frameworks established by industry leaders such as the BSA's AI Bias Risk Management Framework you just heard Victoria speak about. Working together with industry and like-minded countries will help create common standards and develop a global approach to AI that is needed in today's digital world. Adobe stands ready to participate in the creation of these industry codes of conduct to ensure AI is developed in accordance with our shared values while still helping AI realise its immense potential to change the world for better. Thank you for your attention.

1-034-0000
Pilar del Castillo Vera (PPE). – I would like before, of course to thank the two speakers for their contributions – we hope we still can have the third one. This is a really interesting debate. It's quite a crucial and critical aspect of AI, the bias problem. The question is how to reduce the bias in the AI system, and the answer is indeed not that simple.

When referring to an AI bias system, we are talking about systematically producing unfair outcomes to specific demographic groups. And to that, we must first define what 'fair' means. Since there is no universal definition of fairness, developers must evaluate the nature of the system they are creating to determine which is the best metric for mitigating potential risks.

What this means is that the complexity of reducing bias is partly due to choosing the most appropriate metric. Nevertheless, even an AI system that has been tested might produce biased results when deployed in real life. In this context, where one size does not fit all (and my question is for all the speakers), how can we evaluate the entire lifecycle of a product or service and what should be the main characteristics of monitoring processes that organisations should implement to validate an AI model?

1-035-0000
Nathalie Nevejans, *expert in AI and robotics law and ethics***.** – This is very complex issue because, as you quite rightly said, it stems from a variety of sources and there are many

problems associated with it, so it's difficult to provide a single answer. As I explained earlier, I think we'll need EU-approved technological solutions. I'm not talking about AI systems; I'm talking about technological solutions that make it possible to detect bias in AI systems. I think this is something very important on which Europe needs to take a decision. That's my answer, but maybe someone else would like to add to it.

1-036-0000

Dana Rao, *Executive Vice-President, General Counsel and Corporate Secretary, Adobe*. – I think that when we think about how you can continue to test AI systems to understand the ongoing impacts of discriminatory results, I think there are a couple of ways we think about it here at Adobe, and I think the first place we think about it is during the creation of the AI feature. And so what we do is we will require our engineers to have a diverse data set and a diverse set of engineers themselves who are making up the AI, because often some of the unfair results are not spotted by a homogeneous group of people. So ensuring that there are diverse people who are looking at the problem ensures that from ground up, the AI is built to spot these issues.

So the first question is they have to build the data set, and then we have this structure at Adobe in place where once you create AI, you have to fill out an AI impact assessment internally and you have to decide whether or not your AI is going to have an impact on our customers. And if it does, then it goes to our review board and then we ask a series of questions about who they've tested it on and is the data set diverse enough: is it large enough to give you an accurate measure of the impact? That's just some of the work we're doing internally. I'm happy to expand on that later.

Externally then, I think the role in public policy is how you test the outcome, and in our view a specific vertical industry should have test data sets that you could then run against a typical AI application, and then you would see the results and they should have results that meet the expectations of that test data set. So it's a testing process that you could use to understand if the AI system that someone is using is performing in accordance with the values that you expect them to have.

1-037-0000

Ricardo Baeza-Yates, *Professor and Director of Research, Institute for Experiential AI, Northeastern University, Silicon Valley*. – I will not use my presentation because I don't want to repeat things that the previous experts have already said, but please read it and thank you for the opportunity to be here.

First, I would like to recap on the different sorts of bias. The main sort is data, but it's very important to remember that it's not only data: for example, it can be the optimisation function of the algorithm. For example, the case of Deliveroo in January this year in Italy is one of the best examples where basically the system was discriminating against a group of people that didn't have any demographic commonality. And basically it was very hard to understand that bias a priori and you need to find it a posteriori.

But second, you have also biases in data that affect religion, for example, in large language models. There is some recent work on violent completions or, for example, a class like in the UK when they tried to basically predict the scores to enter university or, more recently, in the Dutch government that had to resign because of trying to look for fraud on vulnerable populations using different algorithms. And they not only have the problem of who the people you are affecting are, but also where are you trying to use AI and if it's ethical or not.

And many times, as some others have mentioned, the bias is amplified from the data and then we need to find the source of that, because that's not from the data and one source will be the algorithm, but still, as other people said, we have this feedback loop. And one important part of

the feedback loop is not only important in social networks, but it's also, I think, even more relevant in e-commerce, in a different aspect, because every one of us is affected by that kind of bias. Exposure bias is not really the tip of the iceberg, it's just the top of the ice cube – the top ice cube in the icicle. Because you are seeing, for example, three recommendations from a thousand possibilities, and that creates exposure bias, popularity bias and then the long tail of items and producers gets affected. So I think also another area for regulation should be e-commerce, not only social networks.

Now, because the main topic of the panel is mitigation, let me discuss a few things about mitigation first. Some people believe that the industry can be self-regulated, but we have seen that even having the right business ethics doesn't work. Even when people have ethical codes, that doesn't work. We don't have enough ethics training. And of course, sometimes motivation is more important than ethics – maybe almost all the time. For example, going to the proposed AI regulation, even if you do a self-risk assessment ex ante, the problem is that you have a conflict of interest between what you want to achieve as a business and the limitations that you should have. Also, for example, knowing the limitations of technology is something that I don't see in all computer scientists that believe that they can do much more that they can. And we have to remember that fairness is an illusory goal, because it's very hard to define what is fairness. And even if you have a recent definition, the systems doesn't have a knowledge of the context. That's the main problem of the application of these tools. Data don't have the context of the usage of the tool. And for example, we don't know exactly what the context of the user or what the cultural context of the country where you're using the tool is. And of course, you don't know, for example, the political context of the usage.

So I don't like it when you regulate technology, because I think regulation should be more like human rights, so they should apply to a problem and the independence of the technology. So I don't want to see, for example, in the future, regulation for quantum computing or other new technologies, because we need to solve the problem itself, and then *(inaudible)* technology. For example, you can solve many cases with just plain algorithms – you don't need to use AI – and that is outside the regulation, for example, that doesn't make too much sense.

Another problem with the current proposal is that it's a risk-based assessment, and risk is basically a continuous variable. It's like race, it's like the colour of your skin. How can we split risk in four different levels if this is something that you continue? So I see the game that will be played by companies by basically trying to diminish the level of risk of their solutions. We have other ways to try to do this, like, for example, *(inaudible)* algorithms. But then we need to think if gaming will be a problem, if an algorithm is completely transparent. We can do audits, certifications, as another person mentioned, but then we have a time limitation. This model changes really fast and will not last too long. We can certify the process more like ISO 9000, but then we need to have stakeholders along the whole process of the development of the system.

1-038-0000

Maria-Manuel Leitão-Marques (S&D). – I'm going to speak in English. In my opinion we need to think about two different types of algorithm bias. First, we have bias that is caused because the data set used to train algorithms is not representative of the whole population. For example, the US Veterans Affairs Hospital experimented with an AI model to help predict sudden decline in their patients' kidney function, but women were under-represented in the data used to train their algorithm. And because of this, the model performed worse for them than it did for men. And there are many other examples in the literature.

But second, there are structural inequalities as well as discrimination in our societies, which create algorithmic biases even if the data is fully representative. For example, there are cases of women having lower limits on their credit cards than their husbands, despite having better credit

scores. One way to explain this is because women's and men's consumption patterns differ. This shows up in the data, and they are *(inaudible)*. Therefore, even if data is representative, individual women are still discriminated against because they are women.

Having said this, I want to ask – especially Nathalie, but maybe others of our speakers (very interesting) – I'll ask them: what instruments, what tools can we use to tackle this bias against women and other minorities? Should we have different instruments to tackle the two different types of bias that I have referred to and outlined?

And the AI Act (the proposal, of course) requires that data sets used for high-risk AI systems are relevant, representative, free of errors and complete. And I would like to ask you: is this feasible with the instruments that we currently have available to us?

1-039-0000
Nathalie Nevejans, *expert in AI and robotics law and ethics***.** – I shall answer the first question, which is very interesting – the second question is too, but the first question is of direct concern to me as a lawyer. We are indeed realising that bias in AI is actually just a reflection of the bias that already exists in human society – so it's our bias. So to combat bias in AI we have to start with combating our own bias. And we already have the tools to do that, as has previously been made clear. There are strict EU rules against all forms of discrimination. So when it comes to human bias, we can address it by bringing these legal tools into play. We have the criminal penalties in place to punish discrimination, racism and so on. After that, we need to move to the next level and track bias in AI that stems from multiple sources, has various consequences and is more difficult to detect. That will be a completely different matter.

1-040-0000
Karen Melchior (Renew). – Thank you very much for the experts. I have a couple of questions regarding the tests of outcomes, which was recommended by Dana Rao. How can we test the outcomes if we proceed with the tests of the data sets as you recommended? And also, what would the panellists recommend in regards to regulating the use of AI via specific technical standards for the use of AI or recommended best practices for algorithm design choices?

Because I think the title of this panel is How to Avoid Bias in AI, but we will have bias. The question is how we discover it and how we work against this. And so it's important not just to look at the data but looking at the design of the artificial intelligence. So I would like to hear the panellists' replies to that. And I think this goes out to Dana Rao and also to Mr Baeza-Yates.

1-041-0000
Dana Rao, *Executive Vice-President, General Counsel and Corporate Secretary, Adobe***.** – Thank you for the question. In the tested outcomes in the way we've been proposing it, as an analogy to help understand what we're talking about – and Adobe is not in this field – but if we were in the field of administering loans, for example, you would want a data set, an industry data set. And I say that because they have the data, so you would need the data from the industry, and you wanted to have a demographic profile that you believe is representative of the society in which this vertical is taking place. And you would have a series of outcomes that you would expect to see given that population.

And then you would say, for example – let's use men and women – if there's a certain amount of men in there and a certain amount of women in there and the AI algorithm is assigning credit scores, you would say that, if I gave you this data set, then on average the credit scores for the men and women should be the same given other objective indices.

That would be a test data set that you could then hand to a company in this business and say, hey, we've seen this AI that you've built. We've gotten complaints from citizens that feel that your loan administration is unfair. Please run it against this data set, and we want to see if the

output of your AI correlates with the data set in our expected outcome. So you would be able to see what happened.

One key part of the way I've framed this issue and talking about testing is this is a very complicated technical problem, to remove bias. And as I started off by saying, it's not even clear that this is an intended goal, because the bias in the data set is typically what is interesting and what you're trying to learn.

So what we keep talking about here is the ability to test the outcomes, because the goal of creating traceability and the other technical measures which are laudable are not very technically feasible, and in order to get to that point, you're going to stop innovation. And that may be the right answer for high-risk AI use cases, but it may be the wrong answer for the many companies like Adobe, which have their own self-regulation and are doing things like, for example, suggesting a font for your next novel because we know what kind of novel you're writing. There's no bias implied there. That's why you want to have the multiple categories.

But testing the outcome allows everyone to create a system that works in accordance with the risk that they have or the risk that they propose. So that's how we think about how the database would work. I do think an industry-led solution is the critical way to go about it, because the industry has the data. The data comes from the people interacting with the software or the hardware, and that's the only way to get this data. So I do think the industry-led approach is critical for this to be feasible.

1-042-0000
Ricardo Baeza-Yates, *Professor and Director of Research, Institute for Experiential AI, Northeastern University, Silicon Valley***.** – Sorry, I disagree with Mr Rao. I think that it's impossible to do that. Basically, there are biases that will never be known in the input data: we have intersectionality, we have unknown biases, we even have issues on reference values for biases because we don't have, for example, social agreements in all of them. The same reason if you want to deal with the bias in the algorithms, like basically doing a model that takes care of bias, then you need to know what is the specific bias that is done in search engines, for example, with click position bias. But you need to know exactly what you're dealing with. And of course, if you try to deal with, for example, mitigating the bias in the output, you already lost too much information. So I think that we should regulate like we have done, for example, with food, with drugs, and that regulation should be independent of technology. So for example, let's say hiring, which is a problem where I see not only bias issues but I see also phrenology issues like were mentioned before, like, for example, trying to detect things of your personality based on videos. Then we need to regulate that independently of the technology. For example, what is the right way to hire people?

1-043-0000
Kim Van Sparrentak (Verts/ALE). – Recent shocking reports from Amnesty International and Human Rights Watch have shown us various automated systems being used in violation of fundamental rights across the EU, specifically in social security systems. And notably in the Netherlands, where I'm from, a child benefit scandal has ruined thousands of lives. Parents were flagged as potential fraudsters and were summoned to pay back all their child benefits ever received. And these systems were especially programmed to flag non-Dutch people as a risk factor. And last week, Dutch newspapers revealed that the Dutch tax authorities, through automated decision-making, compiled secret blacklists strongly biased against people with low incomes, and again, without people knowing and let alone with accountability or redress options. And just as Mr O'Flaherty and Mr Leufer pointed out during a previous panel, debiasing is not always enough and not simply the solution.

I would like to address my questions to Nathalie Nevejans and Ricardo Baeza-Yates. First of all, do you think there are situations where we should be more critical than simply relying on

debiasing and on diverse and high-quality data sets? Should we in some situations look further, such as assessing the impacts on other fundamental rights before AI is deployed in a certain situation? Or should we perhaps not rely on algorithms in certain situations at all but require a human to be able to explain exactly why certain decisions are made and why they are justified, rather than an algorithm with a human somewhere in the loop or even human oversight?

1-044-0000
Nathalie Nevejans, *expert in AI and robotics law and ethics***.** – You asked several questions in one, there, and they were all very interesting.

We do absolutely have to be aware at this point that there are AI systems in existence that are very well designed, and the end-user – a company, for example – which is looking to recruit, is going to ask for one criterion or another to be changed so that the company concerned can recruit the kind of people it wants to recruit, thereby excluding others. So bias can have a whole range of origins.

To move on to the use of humans, that is clearly one of the solutions. And the use of humans is indeed a very important factor in the text of the proposal for a regulation, and something we really need to focus on as soon as there's a decision that's going to be taken by a machine – particularly in the case of decisions concerning loans, social security arrangements or social care, for example. So that's really important.

Similarly, there are cases where it's impossible to solve the problem of bias, especially where we're faced with very inexplicable AI. And here I don't think there should be any hesitation. If we use a sandbox to test that AI in situ and – despite the corrections that have been made – we realise it's still biased, unfair and discriminates against certain categories of people, then there shouldn't be any hesitation: it shouldn't go onto the market; we should simply say that AI is not perhaps the best solution in that case.

To conclude, I would make it clear that the AI we're talking about here is, more often than not, machine learning. That said, of course, if we use symbolic AI – so that's expert systems – in this case the way bias acts is much easier to solve because everything is more easily or even completely explicable.

1-045-0000
Gilles Lebreton (ID). – I should like to thank the two speakers, and in particular Ms Nevejans, whose remarks I was interested to hear, especially her proposal to bring in certification for technological solutions designed to detect bias.

I have three questions for you, Ms Nevejans.

First, you said that the origin of bias was always human. But don't you think that artificial intelligence has now evolved to such an extent that it is capable of creating its own selection criteria, and therefore that bias of artificial origin is also a possibility?

Second, you said that some bias is fair, and you gave an example of bias that is geared towards giving women an advantage, for example in the context of recruitment. But isn't that dangerous? Shouldn't we be tracking all forms of bias in artificial intelligence, even if we then – of course – leave it up to humans to decide what the distinctions are, if necessary?

And third, you suggested that research on bias should be carried out on all artificial intelligence systems, and not only high-risk ones. I would state that I myself have focused on high-risk systems because I am the author of a report on artificial intelligence in the military and in sovereign affairs, where the risks are very, very high, especially as regards justice and health.

But why are you suggesting that? Can you give an example of a risk stemming from bias in an artificial-intelligence sector that is not high-risk? Thank you for answering my questions.

1-046-0000
Nathalie Nevejans, *expert in AI and robotics law and ethics***.** – Thank you very much for those very specific questions.

As far as your first question is concerned, you are quite right. Bias is indeed very often human in origin, but because it uses correlations, it is possible for AI to obtain biased results, and it is indeed the machine itself that will do so. So you are quite right to point that out, to temper my perhaps rather expansive comments on that point.

On your second question concerning tracking bias, it's basically a question I'm asking you. It's for legislators, I think, to answer that question, because in some cases – recruitment, for example – there are 500 applications from men and 10 from women. So the question is, do we re-establish that inequality? That's the question. I don't have the answer. And in some cases it could well be that women are less represented in employment. We could envisage influencing the system to restore equality and fairness. That's something that I think should be debated by researchers and politicians, because it's an important issue.

On the last point, that's also a question I've been wondering about. I don't necessarily have an example to give you, but we could think about chatbots. France's national digital council has, for example, published a very detailed report on chatbots and the risks they pose to people. And where video games are concerned, for example, are there not cases that are not high-risk, and yet in one way or another discrimination does ultimately occur? That question needs to be asked and evaluated. So obviously I don't necessarily have the answer there, but I do have the question, and it's a question we need to be asking the researchers and politicians about. I think there is some research to be done there.

1-047-0000
Elena Kountoura (The Left). – Mr President, the private and public sectors are increasingly seeking to automate decision-making processes by means of artificial intelligence systems and machine learning algorithms. Algorithms can be used to formulate action plans based on predictions regarding matters ranging from our political leanings and patterns of consumption to criminal tendencies, health, insurance and debt default. However, they are not objective. Furthermore, we are aware that a number of systems incorporate algorithmic bias that generally targets the most vulnerable populations, thereby exacerbating inequalities and discrimination and highlighting the need to ensure transparency and accountability, both of which are essential for the proper functioning of algorithms and the protection of basic individual rights. The main problem, however, is the difficulty of detecting such bias, given the inbuilt opacity of certain AI-based systems, which is a concomitant of their ability to learn from experience and improve their performance accordingly. This can create a 'black box effect' that makes the decision-making process extremely difficult to follow. Given that the resulting decisions frequently have a significant social impact, algorithmic systems must designed and used in a responsible manner vis-à-vis the public. Algorithmic accountability should include the obligation to report, explain or justify algorithmic decision-making and to mitigate any adverse social impacts or potential damage. To this end, technical and practical measures should be taken to ensure transparency and non-discrimination and details should be provided regarding the workings of mass data analysis, thereby helping individuals to understand and keep track of the decisions affecting them. I should therefore like to ask Mr Ricardo Baeza Yates how decision-making accountability can be ensured with the use of algorithms, notwithstanding their 'black box effect'.

1-048-0000
Ricardo Baeza-Yates, *Professor and Director of Research, Institute for Experiential AI, Northeastern University, Silicon Valley***.** – Thank you for the question. It's a hard question. For

example, we have the Uber case, where in the end the person that was declared at fault was the driver – the safe driver that was inside the car in Arizona, which was not fair. So we need to at least regulate who is to blame if something happens.

For example, I believe it to be the one that provides the system with the product – in that case it was the Uber and not the person inside. Of course, that person can be part of the blame if they *(inaudible)* up, but the main person accountable should be the company or the government, for example, in the case of the Dutch Government, basically providing the solution. And for that, we need to check, as I said in the chat, the competence, because we have not only political competence, but also we have the technical competence and then the ethical competence. Sometimes we don't ask the right questions before starting to use AI.

1-049-0000
Adriana Maldonado López (S&D). – Firstly, and very briefly as we have gone beyond the time limit, I would like to thank the speakers

It is clear that artificial intelligence is created by humans and from human-generated data, which means that this training data already presents, or can present, bias. This is what we have been talking about all afternoon. The concept of what is fair or what we consider a bias or not is a very interesting debate to me. It is also a debate that we need to go into in detail.

We must be able to develop new technologies with artificial intelligence, so that they can be one of our allies in eliminating those biases. I think that is the major challenge of our time: how to use and develop new applications and new artificial intelligence to help us, ascertaining what concrete measures we can achieve with artificial intelligence to eliminate all these biases and with far more fairness.

I am sure the speakers and other colleagues are familiar with this – many examples have been given this afternoon – but I would also like to give an example of how bias can interfere with our lives today.

Amazon used to use a self-learning tool to process job applications and, at a certain point, the system learned to prefer men and, therefore, discarded all applications from certain universities that have a higher proportion of female students. We cannot allow this in today's world; we must use artificial intelligence in a way that eliminates these biases.

I know it is a difficult challenge, but I would also like – and this has already been done throughout the afternoon – to go into a little bit more depth on how we can develop these elements of artificial intelligence to eliminate these biases that we have today.

1-050-0000
Chair. – So Professor Yates, try in one minute.

1-051-0000
Adriana Maldonado López (S&D). – The question is for our visitor from Pompeu Fabra University, Ricardo Baeza.

1-052-0000
Ricardo Baeza-Yates, *Professor and Director of Research, Institute for Experiential AI, Northeastern University, Silicon Valley***.** – The Amazon example was never put in place – that was good. Maybe we're focusing too much on trying to solve the problem with technology. I don't think bias can be solved by technology. I think this is one fallacy, that some people believe that – for some people what's called technological humanism – is more of the same. So basically, technology cannot solve problems like racism or sexism. But for example, I think society can change that with affirmative actions. Let me tell you a very short example, that sometimes the perception that you're doing something is better than the action itself. For example, in Chile, my native country, they decided to give the five last percent only to women

in engineering. And what happened after five years, they were expecting to increase the percentage of women in engineering by five percent, they found it was 12 percent. And why? Because more women applied because they saw that they were doing an action on something that is a problem, and then the impact of the perception of the action was much larger than the action itself. So sometimes we need just little movements in the right direction and then more people will take the same direction, and that would be my recommendation.

Closing remarks

1-053-0000
Chair. – Thank you very much, and I would like to thank all the three speakers, all the three panellists for their valuable contributions, and we have two more minutes, which I would like to thank the interpreters for, and I will give the floor to our rapporteur, Mr Axel Voss, for trying to wrap up in two or three minutes this very rich discussion and give some concluding remarks. Mr Voss, you have the floor.

1-054-0000
Axel Voss (PPE). – Mr Chairman, the discussion today basically dealt with the following questions: What are distortions, what are biases in algorithms? What is fairness and how can we define it?

We need to deal with the question: What is bias, what is fairness, what are prejudices in algorithms, and what are their main impacts? More specifically, we need to focus on high-risk models, in relation to which it must be said that sometimes there are business models possibly targeting only specific groups, but these models might no longer be high-risk based. What we can do is to try and eliminate the obvious issues.

But we have also been told that this cannot be fully achieved, nor can it always be done in advance, and that there may also be both deliberate and random things. How can this be identified in advance? What is required to minimise biases in advance, or how can we ultimately combat bias?

In essence, the question is this: How do we build trust among our citizens? This can be done proactively or reactively. Owing to time constraints, I will only mention a few key aspects: We should basically avoid feedback loops. To do so, we need so-called sandboxes. We need tested outcomes. We also need the checks and balances mechanism. We need compliance structures in companies, but also a governance structure for control, to allow for appropriate indexing and transparency on how to eliminate such things when they occur, and to possibly provide the appropriate technical support for optimisation models. International cooperation was mentioned, just as regulatory approaches to also monitor the life cycle of algorithms, to monitor the development and also the various stages, risk assessment, technical solutions, explainability, transparency, remedy mechanism and quality data. These are all points that can serve as a guide.

1-055-0000
Chair. – Thank you very much, Axel, for your words. Many thanks to the interpreters for bearing with us – over the time limit – to wrap this hearing of ours, the last hearing of AIDA for 2021. Many thanks to our audience for staying with us. Many thanks to the Members for the interesting Q&A. We now close the session.

(The hearing closed at 18.54)