HOUSE OF LORDS

Select Committee on Communications and Digital

Corrected oral evidence: Freedom of expression online

Tuesday 23 February 2021

4.10 pm

Watch the meeting

Members present: Lord Gilbert of Panteg (The Chair); Baroness Bull; Baroness Buscombe; Viscount Colville of Culross; Baroness Featherstone; Baroness Grender; Lord Griffiths of Burry Port; Lord McInnes of Kilwinning; Lord Stevenson of Balmacara; Lord Vaizey of Didcot; The Lord Bishop of Worcester.

Evidence Session No. 12 Heard in Public Questions 106 - 113

Witnesses

I: Professor Carissa Véliz, Associate Professor, Faculty of Philosophy, Institute for Ethics in AI, Hertford College, University of Oxford; Professor Sandra Wachter, Associate Professor, Oxford Internet Institute, University of Oxford.

USE OF THE TRANSCRIPT

This is a corrected transcript of evidence taken in public and webcast on www.parliamentlive.tv.

Examination of witnesses

Professor Carissa Véliz and Professor Sandra Wachter.

Q106 **The Chair:** Welcome back. We have some new witnesses, Professor Sandra Wachter and Professor Carissa Véliz. We are going to talk, in this session, about artificial intelligence and the aspects of artificial intelligence relating to freedom of expression online. The session will be broadcast online today and a transcript will be taken. Professor Wachter is associate professor at the Oxford Internet Institute and Professor Carissa Véliz is associate professor at the Faculty of Philosophy and the Institute for Ethics in AI at the University of Oxford. Thank you both very much for joining us today and giving us the benefit of your expertise. Can I ask you to introduce yourselves, tell us about the institutions you come from and then give us an overview of your perspective on freedom of expression online from the point of view of your areas of expertise?

Professor Sandra Wachter: I am an associate professor at the Oxford Internet Institute. It is an interdisciplinary research department within the University of Oxford. We have various scholars working on the governance issues surrounding technology. I am a lawyer by background and am interested in the legal and ethical implications of emerging technologies. At the moment, I am working very strongly on issues of AI and machine learning, but I also have expertise on internet regulations and human rights online. From a legal perspective, free speech is probably one of the most important human rights that we have in a democracy. It is definitely both a condition and a result of democracy.

It is very important to keep in mind how many faces freedom of expression has. Many of us will just think about communicating with friends, for example, or writing something down on a piece of paper, but it is much broader than that. It includes, for example, journalistic expression, the free press, freedom of art and freedom of science. Freedom of religion is a form of expression. The rights to assembly, protest and unionise are forms of expression. Lastly, the right to vote is a form of expression, so the nuances are very wide ranging and I cannot overstate how important it is. It is really a founding stone of our democracy.

Professor Carissa Véliz: Thank you very much for inviting me. It is such a privilege to be here. I am an associate professor at the Institute for Ethics in AI at the University of Oxford. The institute is very new. We started just in October, which speaks to our times and the great necessity to think more carefully about ethics. We are based in the Faculty of Philosophy and some of my colleagues work on democracy, the role of judges in the context of algorithms, language and many other topics.

On free speech, I want to convey at least three messages. Toxic speech online is fuelled by a toxic business model online. If we got rid of the business model and the incentives that kindle the worst kinds of speech, social media platforms would be much better places in which to interact and we would have far fewer problems and a lot less need to curtail speech. The toxic business model depends on the exploitation of personal data.

Essentially, social media has an incentive to make people engage as much as possible mainly because it wants to collect more and more personal data. It turns out that the most engaging content is really toxic content, from fake news to hate speech. If personal data trades were banned, which is what I argue for in my book, *Privacy Is Power*, we would not have that toxic incentive and the landscape of social media would radically change. We should not lose that focus.

I wrote a paper a couple of years ago called “Online Masquerade”, in which I argue that we have lost the practice and the value of pseudonymity, and this is a huge mistake. Most writing before the Renaissance was anonymous. In philosophy, for instance, John Locke’s *Two Treatises of Government*, Hamilton, Madison and Jay’s *Federalist Papers*, Marx’s communist manifesto, Kierkegaard’s work and many others were published either anonymously or with a pseudonym. If we had not given them that chance, those works might never have seen the light of day. Today, we have the balance wrong on the internet, I argue.

The problem with anonymity is interesting. We want desirable speech to have no cost. We want people to say things that are valuable for society and not experience any kind of negative consequences. At the same time, we want negative speech to have consequences. We want people who use hate speech and other destructive speech to face repercussions. If we have full anonymity or full identifiability, we cannot meet those goals. If we have full anonymity, the bad guys will not face any consequences, but if we have full identifiability very valuable speech will be crushed out. Right now, trolling is one of the main threats to freedom of speech online. It is pushing women out of politics. Academics are self-censoring out of fear. It is pushing activists and other political dissenters out of the public sphere. This is a huge problem.

I argue that we can use pseudonyms to regulate the cost of speech. Imagine that there was an institution that gave out two or three pseudonyms to every person, and they would act as fiduciaries to the link between the real name and identity of a person and their pseudonym. You have two or three pseudonyms and you can use them throughout your life as you want. You might want to use one of them for years on many platforms. You might want to use one for Twitter and another for Facebook. If you use hate speech or if you abuse speech in some way, you can have a penalty or a warning, or you can have your pseudonym taken away from you. If you run out of pseudonyms, you have to either not engage online or engage with your own identity, which is much riskier.

In this way, we could regulate speech so that people could speak their minds and would not face very bad consequences, but if they sent, for example, a death threat, this organisation that gives out pseudonyms would have the ability to reveal the name to the police. That is a system that we should take seriously. We should think more about how we can use pseudonymity to protect people and, at the same time, allow for free speech.

It is time to act. It is time for regulators to spring into action. Efforts cannot end in having conversations and creating reports. It is really urgent. Our democracies are in danger and the time has come to regulate big tech. We have regulated all other kinds of industry: railways, cars, aeroplanes, drugs and food. There is no reason why we cannot regulate big tech.

In this context, diplomacy will be very important. The danger of having each country come up with its own rules is that non-democratic Governments will come up with rules that abuse those powers and create very bad repercussions for political activists. We are already seeing this in much of the world, so we need to create an alliance of democracies in which we come up with the rules. We should not let private companies such as Twitter and Facebook come up with the rules of engagement, because they are private companies and do not have the public interest at heart.

The Chair: Can I pick you up on this idea of pseudonyms and how it would work? Who would decide that someone had said something that reached the threshold for being named, shamed and identified? It would not be the tech companies, I guess. What body of people would decide that? In the case of death threats, it is obvious. That is illegal in any circumstances, but who would decide that the speech is so hateful that the person should be named and shamed? Would a system of pseudonyms work in an offline environment, as in the case of academics who fear that, in their university, they cannot express themselves because of pressure from colleagues?

Professor Carissa Véliz: I do not claim to have all the answers. This is something that I am thinking through and we need more people to think through. In the case of academics, Francesca Minerva, who works at the University of Ghent,[1] has proposed a system in which, as academics, we could publish online under a pseudonym and then we would have a system that connects all the universities. Let us say that you are going to apply for a job. Then the university that is going to employ you gets a one-time code to verify which articles you have published, so that could be something to think about.

Regarding your first question, ideally, it would be an international organisation, made up of an alliance of democratic countries that come up with the right rules. I argue that, for crimes, people should lose their pseudonymity and the police get their identity. In many cases, what counts as very bad speech is not clear and we should have some kind of leeway. That is why this system creates a buffer. If we do not allow people to say things that are unpopular, society will never change. At some point, it was really unpopular to say that women should have equal rights and that slavery should not exist. If we do not allow that kind of speech to enter the public realm, we will never change.

[1] Amended by witness: Should read "University of Milan"

An example I give in my paper is that, at the moment, it is controversial to ask whether animals should have the same rights as, say, toddlers because they have the same kind of abilities. Some people get very offended by this kind of speech. Say you have a pseudonym for three years and you have thousands of followers on Twitter. If you voice this opinion, it has consequences. You will have to face your peers and defend your argument, but it is not so bad. If your speech does not accuse anyone of anything untrue, incite violence or anything like that, you shouldn’t face such grave consequences that would amount to having your life completely wrecked and being unable to ask for a job because, when people search for your name, horrible things come out. You should not have your life destroyed because you have an unpopular opinion.

At the most, if this organisation decides that your speech is so questionable that you should lose a pseudonym, you still have one left with which you can be more careful. This system creates second chances and some leeway to decide what counts as really hateful speech. Much of it will depend on trying it out and putting it into practice.

The Chair: That is very interesting. I find it slightly depressing that the state of academic freedom, as you describe, is such that people need pseudonyms to express themselves in an academic environment. None the less, you describe a process that could work. You described it more specifically in relation to online; I accept that.

Q107 **Baroness Buscombe:** You both gave excellent introductions. The need for a pseudonym or something else has probably existed for a fair number of years now, which is a sad reflection on so-called technological progress. My question is drilling down into the detail on this. Is large-scale automated content moderation compatible with freedom of expression? We know that platforms are increasingly reliant on automated processes in content moderation. For example, can this lead to making contextual errors and miss the significance of content? Is there a risk of overcensorship?

Professor Sandra Wachter: It is a critical question. To some extent, an automated algorithmic moderation system could work in practice, but only if you have in mind a very clear target. A very simple example would be procedures when it comes to copyright. If you know that a certain type of work is copyright-protected, it will be very easy for an algorithm to sift through the internet, compare pixels and take it down. There, you have a clearly defined goal in mind.

However, very often, with free speech matters, it is not black and white. It is muddy, grey and contextual. An example would be sarcasm or satire. There is interesting research that shows that both humans and algorithms are able to detect sarcasm in only 60% to 70% of cases. Both are quite bad at understanding the underlying subtleties of human language, so there is a chance that algorithms are not very good at detecting this.

It is very unclear what "harmful" means. We have discussed that for centuries. A lot of things that used to be seen as harmful or obscene no longer are. People's minds are changing on that. An algorithm is not very good at detecting this. We have to think about false positives and false negatives. A false negative means that something is left up there that should be taken down. You can game the system just by using odd spelling that the algorithm does not recognise—for example, a zero instead of an O. The algorithm will not be able to detect it. In the same way, it could overpolice it and think that something is harmful.

An interesting example came up a couple of days ago, where content was taken down because it had the words "black", "white" and "attack" in it, but it was actually about a chess game. The subtleties are very important, and human language and cultures are framed around subtleties, so I would be very worried about saying, "Yes, algorithms can absolutely do that". They might be able to do it for a very clear and narrow range of things, but, as soon as we talk about fuzzy concepts, I would be very worried about that.

Professor Carissa Véliz: I completely agree with Sandra. In theory, you could have an algorithm that is so good that it identifies hate speech only. In that case, it would not be a threat to freedom of expression because nobody has a right to hate speech. The concern is that algorithms might never be good enough to understand the subtleties of language. Language, as Sandra said, is very contextual. It has a lot to do with the offline world and algorithms are not embodied beings with empathy that can understand what is going on, so I have a lot of scepticism as to whether they will work very well.

That is one reason to prefer other kinds of platforms, for instance those in which there are smaller private groups. Human beings are a lot better at moderating each other in very small groups than in huge groups online, in which the moderators may be from a different culture or may be algorithms with the help of human beings, who are always overwhelmed with work.

Baroness Buscombe: That is helpful. We are all learning, through living our lives on Zoom et cetera, the difference between the large and the small, the nuances and what we are missing by not being physically together. The expectation that machinery can do a better job than we can is quite extraordinary.

Q108 **Viscount Colville of Culross:** I would like to take on Baroness Buscombe's question about automated content moderation. There are studies that show that AI content moderation disadvantages certain groups. They found that AI models processing for hate speech are one and a half times more likely to flag tweets as offensive when written by African Americans, while Urdu and Arabic content suffers even greater flagging. How concerned are you about the effect this has on discriminating against certain groups?

Professor Sandra Wachter: When we talk about algorithmic systems, bias is one of the most pressing things we have to address. It is not just

African Americans. The problem with AI is that the oppression of all the groups that are already disadvantaged is exacerbated by the use of algorithmic decision-making. That is a really big problem. The reasons are very diverse, unfortunately, so it is hard to put your finger on that one thing that does it, but it has to do with the bias of the people who label the data. It has to do with the lack of context that, for example, content moderators get, where they do not understand the cultural or gender differences of the content they are looking at. Very often, different datasets are combined, which exacerbates this. That is a very big problem.

What can we do about it? Again, there is no one solution that will do the trick. We have to tackle this on various ends. From a legal perspective, we have to make sure that we have appeal mechanisms in place where there is a human in the loop at some point who can look at the content. This in itself raises ethical questions, because being confronted with, potentially, extremely harmful content on a daily basis is not necessarily an easy job. We need to keep this in mind, and think about training opportunities and support systems for people doing that job.

From a governance perspective, you can do things on a tech level that would be very helpful. One would be so-called tech data provenance, which means you have more information on where the data comes from, who collected it and its purpose. You can think of it as a nutrition label that lists the ingredients of what you are dealing with. That is a transparency tool. It is not going to solve the underlying problem, but at least you understand what you are dealing with a little better.

In the last couple of years, I have worked on with my team on bias tests and explainability tools. With my colleagues Brent Mittelstadt and Chris Russell, I have published several papers on how we can make sure that, if algorithms are making decisions on what content gets taken down and what gets left up, the outcome is fairly distributed among certain groups. There are ways to test for that. We wrote a paper called "Why Fairness Cannot be Automated",[2] which I am very happy to share, that shows what mechanisms are possible that also take into consideration the legal notion of fairness. There are tests that could tell you whether you are being fair.

On explainability tools, you might wonder why your content was taken down or left up. We have worked on what are called counterfactual explanations,[3] to tell you why your content was flagged: "Your content was taken down because it contained these words" or "It was taken down because those pixels were in it". It would give you the opportunity to challenge that, if you felt that it was unfair, out of context or misread.

Both those tools are quite easy to implement and not very burdensome. Both Google and Amazon have recently taken up our work on counterfactual explanations and bias tests, and implemented it in their work. From a tech perspective, you can do things to battle this very

[2] https://arxiv.org/abs/2005.05906

[3] https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

important problem, but a lot of avenues have to be explored at the same time.

Professor Carissa Véliz: I largely agree with Sandra. We have to constantly audit algorithms, because data is changing all the time. It is not enough to make sure that an algorithm is working well at one particular time or with one particular case. Furthermore, I find it astonishing that we are allowing algorithms to be let loose into the world without having been rigorously tested. Essentially, we are treating the population as guinea pigs, and that strikes me as very problematic.

Look at the history of medical ethics. In the 1940s and 1950s, or before then, you might have gone into hospital and been enrolled in an experiment without knowing about it, without being given any kind of informed consent or receiving any kind of compensation. Medical ethics changed that, and digital ethics has to do exactly the same thing. It is not okay to treat the general population as a testing ground. Algorithms can have as bad an effect as the most powerful drug you can imagine, yet we would never allow a drug to go out into the world without randomised controlled trials. Not even in an emergency situation such as the coronavirus pandemic did we allow vaccines to get out without really trialling them first. In the same way, we should not allow algorithms to have a significant impact on people's lives without having gone through randomised controlled trials.

Viscount Colville of Culross: Thank you very much indeed. That is fascinating.

Q109 **Baroness Featherstone:** Algorithms seem to be the enemy. They seem dangerous, from what you are saying. One rationale behind this, which has come up in previous meetings, is the lack of transparency about how they rank content. They seem to have attracted a perceived unfairness in making stories from some sources more prominent than others. They have a commercial impact. If you are put down the rankings, you are going to lose money, but, if particular sources or campaigning websites suffer reduced positions, that could also impact freedom of expression.

Although we were told by one platform that it ranks news sources by relevance, prominence, authoritativeness, freshness, location and usability, we had no detail about how platforms define those things. Google told the Committee that algorithmic transparency, with disclosure of raw code and data, would raise huge risks. Are platforms sufficiently transparent about how their algorithms rank content in search results and social media feeds?

Professor Carissa Véliz: No, they are not transparent enough. Not even the best experts I have ever met can give a clear explanation of how algorithms are ranking stories and content. More concerningly, companies have an interest in ranking content according to what is most profitable for them—of course: they are companies. That seems incredibly problematic, because it turns out that what is most profitable is the most toxic speech you can possibly imagine. That is really good money because it engages people. It makes people angry, pit themselves against each other and engage endlessly online. All the while, platforms collect their personal data and use it to fuel that system, in which we get more personalised content. It is just a vicious cycle.

One way to break it is ending the trade in personal data. That is very important. When we talk about transparency, it is tricky to know what kind of transparency is meaningful. If these companies gave us the code, we would not be able to do much with it, even if we were super-experienced programmers. The code will be millions of lines long and will be obscure, so transparency of the code, although in some cases might help, is not enough. Transparency of the objectives of the companies is really important. A company designs an algorithm with an objective in mind. It tells the algorithm to do something. To have transparency about what exactly is being given preference is as important as other kinds of transparency.

Baroness Featherstone: That is a good point.

Professor Sandra Wachter: We know that one of the business incentives of tech companies is to keep you on the platform and, unfortunately, what keeps us interested is gossip, scandalous talk and toxic behaviour. That is just human nature, so it is clear that the business model has emerged from that. It is also true that companies such as Google, Facebook and Twitter have made attempts to be more transparent, telling the users what is being inferred about them and what they are learning about them. This is a good first step in the right direction, but it is very clear that there is more going on behind closed doors than they make us believe.

What is really important, and I cannot stress this enough, is that, from a legal perspective, the ideas of fairness, justice and equality are based on the idea of comparison. Am I treated fairly? I can only assess this if I look at how you are being treated. This element of comparison is completely eroded in the online world. For example, if I go to Amazon and I am offered a price, I do not know what price you are seeing. If both of us were at Tesco, we would be able to see the same price. If somebody was to offer you a different price, I could raise a complaint, but I do not know if I am getting the best product and the best price any more.

Similarly, if I am, for example, looking for a job or a loan, an algorithm can infer very sensitive details about me, such as my sexual orientation, my gender and my political beliefs, and use this information to tailor the world to me. I might be looking for a job and be filtered out before I ever see the job advertisement. I will never complain because I just do not know. In the offline world, we would see, for example, the paper. Open up the *Guardian* and we can both see the job ad. This comparison element is eroded and people are being siloed in little pockets, where they do not know what the whole picture is.

That is, from a legal perspective, extremely important to keep in mind, because a lot of legal tools are built on the idea of knowing what is happening around us. I have raised this in two of my papers, one dealing with non-discrimination[4] issues and one called "A Right to Reasonable

Inferences”.[5] I think there is something we can do. I advocate for a right to reasonable inferences. That means that, usually, we have a right to be reasonably assessed. I am very much in favour of moving away from loading the burden on the user to manage their privacy preferences on a daily basis. Nobody can expect that from a human being. Nobody has the time, resources or understanding to do all of this.

We need to give the responsibility to those who collect and profit off the data, and then we collectively have a discussion of what a right to reasonable inferences would mean. It means they have the responsibility to act in our best interests and see it as a collective right. Yes, there is greater need for transparency, so I do understand what a company learns about me and I have the ability to see if I am fairly treated.

The bias test[6] we have developed, which I mentioned before, would allow you to do that. It would be possible to implement those bias tests into business practices and, for example, publish the results of summary statistics, so consumers, regulators and judges could see what kind of bias tests have been done, how often and what the results are. Very often, I will not know if I am being treated fairly. Somebody needs to give me that information. It is possible to do this in non-burdensome ways that do not come into conflict with intellectual property rights.

The Chair: Sadly, we are quite pushed for time, so I will ask you to answer the questions reasonably briefly. In one or two cases, you may want to send us further elaboration afterwards, if you have time. It would be really appreciated in some of these complex questions.

Q110 **Lord McInnes of Kilwinning:** I am quite keen to look at two separate areas where content interfaces with the user and how that is best managed. I will ask each of you a slightly different question. Carissa, you mentioned that the platforms are, after all, businesses. I would like to hear from you how we can balance the user’s right to control the content they see, even if that involves bringing in, for example, third-party applications to the business, and ensuring that the platform remains a viable business. How do you think that control would work and where is that limit?

Sandra, I would like to hear your view on editorial policy of platforms, specifically de-amplification of stories, where the platform makes the judgment. How would your bias test fit into that? How could that be made a transparent process where people could be clear on what the platform was doing?

Professor Carissa Véliz: That is an excellent question. We should question the business model. We should not think that, just because Facebooks wants to earn its money in a particular way, that is acceptable. It might sound radical to end trading personal data because we have become so used to it, but, really, what is extreme is to have a business model that depends on the systematic mass violation of rights.

[4] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3388639
[5] https://journals.library.columbia.edu/index.php/CBLR/article/view/3424
[6] https://arxiv.org/abs/2005.05906

Take Google. In 2013, Google already earned enough money to be one of the biggest companies in the world. There was an article published by *Forbes* that calculated that Google really earned $10 per year from users. Just like we are used to paying for newspapers and Netflix, there is no reason why we should not pay for other things. It might not be very expensive.

I argue in my book, *Privacy Is Power*, that we should implement fiduciary duties. Fiduciary duties are duties that are relevant when there is an asymmetry of power between a professional and a client. Typical cases include doctors and patients, lawyers and clients, and financial advisers and clients. It is a relationship of extreme vulnerability. You are giving something to the professional that is really valuable, your body, your finances, your legal case or your personal data, and the professional can have conflicts of interest. You can imagine your doctor might want to earn more money by performing surgery on you, and they cannot do that. They can act only if it is in your best interest. In the same way, anyone who collects or manages our personal data should have a duty of care.

If you do not want to assume the responsibility to use personal data only for the benefit of the data subjects, you can choose another job, just like it is not enough for a doctor to enjoy cutting up people—you have to accept the duty of care, otherwise you cannot be a doctor. You should not manage personal data if you are not ready to keep it safe and to use it for the benefit of data subjects.

Lord McInnes of Kilwinning: On that point, there could be a situation where the user is quite happy for that content to be managed for them. I guess it comes down to transparency, so that the user can make an informed choice.

Professor Carissa Véliz: That is one thing, but another thing I argue is that privacy is a collective enterprise and a collective concern. For instance, in the Cambridge Analytica example, why do we think that a person has the moral authority to give out their data if that data is going to, first, enable Cambridge Analytica to access the data of many other people who did not consent and, secondly, be used to create a psychological tool that will try to sway elections by manipulating people?

In a sense, my personal data contains all kinds of personal data about other people. My location data contains data about my neighbours. My genetics contains data about my parents, my siblings and so on. I do not have the moral authority to give up certain kinds of personal data. We should have some minimum standards of things that you cannot do with personal data just like, in this society, we do not allow votes to be bought and sold. There are certain things that should not be done to personal data because it is just too risky for society.

Lord McInnes of Kilwinning: Sandra, what is the effect of de-amplification on freedom of expression online?

Professor Sandra Wachter: That is the hardest question of all. From a legal perspective, it is important to keep in mind that almost no human right is absolute. It always has to be counterbalanced with other interests. The first step to do that would be to figure out whether there is a public interest to limit that particular right. We are still at the stage of figuring out how harmful the threat of misinformation is and what effects it has on people. We need a good evidence base before we make those decisions. Once we have that, we can think about various tools to govern that.

I would prefer to have a nuanced approach, because it is such a grey area and we cannot have one solution that fits everything. That can range from labelling certain content as disputed or maybe as wrong, but still giving people the opportunity to make up their own minds, through to blocking a person if they are doing something extremely harmful. We have to allow ourselves that nuance. Just because algorithms operate between ones and zeroes, it does not mean we have to. It is very important to keep in mind that there is nothing wrong with us; there is something wrong with the tech. The tech has to serve our needs. If it does not, we have to go back to the drawing board and come up with technology that is more diverse, like the humans we are.

The bias test and the explanation tools that we have developed cannot give you the right answer of what is right and wrong, and I would never want to have that. We are supposed to grapple with those questions. It is our opportunity and our responsibility to figure that out. We should not shift that burden to technology. That is not what it is supposed to be doing. The tools I have developed are there to make sure that you know that you are doing a good job. Once you have decided what the right path forward is, you have a checking system that makes sure that you are not disadvantaging certain groups. If you do not understand how an algorithm works, you have an ability to investigate further. That is the role of technology. It is not about giving us the solution. It is about enforcing the solution that we have agreed upon.

That comes back to the question of what the right solution is. There are so many interesting examples out there that we should take into consideration. I am greatly in favour of not overburdening the individual. A right to reasonable inferences[7] could be great. I like Jack Balkin and Jonathan Zittrain's work from 2014 on information fiduciaries[8] and fiduciary duties of care. There is a lot of promise there. They have written this up extensively.

I like Brent Mittelstadt's approach involving licensing tools,[9] making sure that the tech community has certain standards similar to other very important people in our society, such as journalists, doctors or lawyers. All those people have very important roles and are trusted with our sensitive information. They should have an ability to show that they adhere to the rules. There are very interesting proposals out there.

[7] https://journals.library.columbia.edu/index.php/CBLR/article/view/3424

[8] https://www.techpolicy.com/Blog/September-2018/Jonathan-Zittrain-and-Jack-Balkin-Propose-Informat.aspx

[9] https://www.nature.com/articles/s42256-019-0114-4

Looking at what people think of responsibilities and then using the tech to enforce it is the right approach.

Q111 **Lord Vaizey of Didcot:** You are clearly capable of auditing algorithms. Should our regulator be given that power in the online harms Bill? Is there a backdoor way of regulating algorithms by making the platforms liable for their rogue behaviour, in the way that they might be liable for the rogue behaviour of an employee?

Professor Sandra Wachter: Audits are extremely important. Ethical audits are the only way to get to grips with the problem. Internal and external audits are very important. I fully agree that not only is making the whole source code publicly available a problem in terms of trade secrets and gaming the system, but it is not entirely useful. I would not understand the full algorithm if somebody gave me the code. Somebody needs to understand and to have access to it. A regulator would be very well suited and this is the perfect way to strike a balance. There does not have to be pure openness, but a regulator or a judge needs to have access and the resources to understand and investigate this. It is going to be tremendously important to do that.

In the same way, we should work on standards that require regular testing inside the company, in the sense that it is not a box-ticking exercise: "I have done my bias test once, so I am done with it". "I have made sure that the data is being collected. It has to go on for a long time." Why should we expect less of algorithms than we would expect of humans? If anything, maybe we should expect more of them. Yes, of course they should be liable for the harm that their algorithms are doing. You are deploying an algorithm because it is time efficient, cost efficient and more accurate. You are reaping the benefits. Well, then, you are also taking the risk that, if something goes wrong, you are liable for it, as with any other technology you use or any other employee you hire.

Professor Carissa Véliz: They should be liable for their algorithms even more than for their employees, because employees are moral agents who make decisions for themselves and do not necessarily obey the company. Algorithms are tools designed by the company for its own benefit and profits, given its own objectives. If they fail, the company is definitely responsible for it. Regulators should have more access to these algorithms and should regularly audit them, not only because algorithms constantly change, particularly machine learning algorithms, but because the data changes all the time. Even if an algorithm is very accurate at time 1, at time 10, after five years, in which practices and people can change, it might no longer be accurate, so we have to audit them very regularly.

One of the problems right now is that regulators do not have the necessary funding, people and expertise to stand against these big tech giants, so we need to fund our regulators much better.

Lord Vaizey of Didcot: I agree with that.

The Chair: Who should pay?

Professor Carissa Véliz: Partly, the big tech companies should pay. They definitely have the money for it.

Lord Vaizey of Didcot: Have licence fees.

Q112 **Lord Stevenson of Balmacara:** Thank you very much indeed for the contributions so far. It has been absolutely fascinating. A recent report from this committee showed that there were about a dozen regulators already with an interest in the digital world. We have been talking about companies that are bigger than many countries, even groups of countries. What sort of regulator would be able to take on this new tech world? Do you think that the duty of care is the right way of going forward in that, because that makes it a joint responsibility between the regulator and the regulated to come up with solutions?

Professor Sandra Wachter: The tech companies are very versatile and their business model is constantly changing. Therefore, the game has changed and, therefore, the rules probably have to change. Tech clearly shows us how it affects so many sectors at the same time. A co-ordinated approach by different regulators will be a very important step forward, involving the ICO from a data perspective, the equality bodies to deal with discrimination issues, and the competition authorities. Siloes are no longer an option because it is such a multifaceted problem.

That might be challenging because we have to learn from each other quite a lot and co-ordinate our work. However, if we do that, we can have a very holistic approach that finds solutions that do not stifle the economy or innovation, but at the same time make sure that human rights are fully protected. There is a way. Usually, that happens when you have different people in a room who are nothing like you because that is how creative energy can work.

That is most important here with AI because the technology is so very different from humans. We have a lot of experience of regulating humans because we understand what motivates them, what reasons they give and how they act. All the tools we have developed are very good for investigating, preventing or punishing them, but humans are not algorithms. They work very differently, so we need to think very creatively in a co-ordinated way to find that sweet spot of using it for the good it can do while, at the same time, preventing it from doing harm.

Professor Carissa Véliz: I suspect that we might need a new regulatory body just for the tech giants. They are so big and co-ordination is so difficult. If we look at the past, typically, when there is a very big industry that needs to be regulated, an agency gets created just for that. That might be the case here.

We have many reasons to engage in diplomacy. The UK should liaise with the United States and Europe. One characteristic of these companies is that they are global, and how they get regulated in one country might shift their behaviour in another, as we are seeing now with Australia. The more countries can come together and agree on policies, the more power they will have against tech giants, which might outpower any single country.

Governments need to focus on avoiding temptations. A concern I have about the UK, and I have written an article about this on the *Guardian*, is the temptation to become a data haven. It is very easy to think that liberalising personal data is an easy way to earn more money and to become an important place where tech companies might want to come and develop products. It is a huge risk. It would be a huge mistake. This model of personal data is undermining equality. You and I are not being treated as equal citizens. We do not pay the same for the same services. We do not wait in line for the same time. Every time we get treated by an institution, we get treated on the basis of our personal data.

Furthermore, it is a huge risk to national security. We have already seen glimpses of it, but the internet has been built in an insecure way, partly to enable the collection of personal data. Part of the temptation is to think that the more personal data we have on everyone, the better chance we have to keep the population safe. It turns out that big data is not the kind of method that works well to prevent things such as terrorism. No matter how much data we had, we would never have thought that somebody might use a pot to create an attack on the Boston Marathon. On the other hand, having this insecure internet is really risking society. For instance, hackers would need to hack only about 10% of electrical appliances and turn them on at the same time to bring the national electrical grid down.

Lord Stevenson of Balmacara: You need to stop now. You are terrifying me.

Professor Carissa Véliz: It is really terrifying.

Lord Stevenson of Balmacara: You have made your point very well. Thank you very much indeed for that.

The Chair: As a point of information, you are talking about reinventing the internet, it seems to me. That will only happen if there is massive global co-operation between nation states, not just at EU level or bilaterally between countries. Is there any sign at all of a significant number of powerful nation states coming to that view and thinking of a way of convening to that end? Do you see anything?

Professor Carissa Véliz: It is definitely a time for optimism. Joe Biden is much more open to collaboration than his predecessor. Europe is very interested in regulating tech, as is Australia, so it is the perfect time to come together.

Q113 **The Lord Bishop of Worcester:** Thank you for a most tremendous session, which has been highly informative, greatly thought-provoking and much more engaging than any Twitter storm I have observed. I have the added consolation of feeling that I am not being manipulated by an algorithm. I have always been a bit wary of algorithms because I felt I did not understand anything about them. Then I was lured into a false sense of security by a previous witness who said, "You just need to think of them as employees". I thought, "Okay, I can get my head round that".

What you have said this afternoon has alarmed me very greatly and I have had to rethink, which just shows how little I really know about how they function. I am wondering how much users really understand about how they function, the answer to which, I suspect, is very little. More importantly, how could public understanding of their decision-making and its ethics be improved?

Professor Sandra Wachter: I fully agree. Understanding algorithms will be extremely important for everybody involved: the end user, regulators and companies. Yes, it is clear that not everybody can be a computer scientist; nor is that a necessity. I sometimes compare it to owning a car. I do not fully understand how the car functions, but I know the basics of it, which I learned in high school. What is important for me is that I know enough that the mechanic and the seller cannot fraud me, and I understand what it can and cannot do. It cannot fly; it does not function underwater. I need to understand when it is broken; it needs to tell me when it is not functioning so I see a mechanic about it. I need to understand the benefits and the risks of it. I do not need to understand everything, but the basics are very important. That means education, from very early on until the late ages of your life. Education should never be over.

Because you do not want to overburden the individual with that, you need to find trustworthy handlers. Using the example of aeroplanes, I do not want to worry, every time I get on one, about whether it is going to crash. I need to trust that those who are creating planes are doing it according to certain standards so I do not have to worry about it all the time.

Now it is time to figure out what those responsibilities should look like. My colleague Brent Mittelstadt is working on questions on licensing. Should we license engineers and think about their responsibilities? As I mentioned, my colleagues Jonathan Zittrain and Jack Balkin have been working, for a long time, on the idea of information fiduciaries and duty of care. I think that a right to reasonable inferences[10] could get us there to figure out what a trustful handler would look like. A combination of these is feasible and would do the trick.

Professor Carissa Véliz: There is something to thinking about algorithms as employees, but one of the problems is that algorithms are not working for you. One question is to ask who the algorithms are working for and what kinds of incentives are there. Visualisation can help us understand algorithms better. I found the new data privacy labels from Apple to be an interesting experiment. They are far from perfect, partly because they depend on people saying what data they collect, so they could lie. When you compare two apps, you do not have to know a lot about personal data, privacy and algorithms to realise that, because Facebook's label is so long, a lot of personal data is being collected, while

[10] https://journals.library.columbia.edu/index.php/CBLR/article/view/3424

another app has a very short list. Assuming they are telling the truth, that tells you something.

Many times, when people say they do not care about privacy or algorithms, I do not think they realise to what extent their lives are ruled by algorithms. One task we have is to make it a lot easier for people to understand and have access to this kind of data, so labels would be an easy way of seeing how much data gets collected. Anytime an algorithm is used to decide something important about a person, that person should know that an algorithm was used and what data that algorithm had access to.

At the moment, when people go to the bank to ask for a loan or to open a bank account and get rejected, they usually just get a rejection. It is really hard to come up with something more. You have to apply to the company that does this work for the bank and it is really complicated. We need to make sure that people can access the kind of data they are being judged on, and that they can contest it and say, "This is not true; this is not fair".

The Chair: Thank you very much indeed, Professor Wachter and Dr Véliz. This has been a very useful session. You have given us a lot of information. You have quoted quite a few sources, so, if you would like to send some of those sources to us, that would be very useful indeed. It will give us something of a reading list. We are not short of reading material, but, if you have time to pull some of those sources together in a quick note for us, the committee would find that very useful indeed. Sadly, that brings us to the end of the meeting.