Transcript of “Contesting Algorithms” featuring Niva Elkin-Koren – October 22, 2019

OK. Let's start. My name is Yochai Benkler and I'm one of the professors here at the law school over at Berkman. We are extremely fortunate-- we are extremely fortunate to have with us today Niva Elkin-Koren, a professor at the Haifa University, Faculty of Law, former dean of the Faculty of Law, founder of both of these organizations, both the Haifa Center for Law and Technology, and now, after she finished with the deanship, the new Center on Cyber, Law and Policy, focused more on various range of threats.

Niva is an old friend and an old colleague. For many years, she has the distinction of always being somewhere between 5 and 10 years ahead of everybody else on most, or at least, on many issues. [? Robot ?] was probably the first significant piece on how copyright intersects with democratic meeting-making and bulletin board services.

And then when we were all very busy talking about non-state, non-market, models, she wrote "The Invisible Hand," telling us that the state was coming back using a variety of levers-- that we today all are familiar with, but at the time, we weren't thinking about-- and in the last few years, has been spending time in several dimensions, looking at the ways in which the weak, in the context of what we often worry about as technology empowers the powerful, can reverse technology in ways that provide new forms of power.

And I see this new paper as very much of a piece with that. And perhaps, Niva's [? presence ?] in the last 25 years suggests that in 5 years, or 10 years, we'll all will be somewhat more optimistic than most of our conversations are today. So Niva, please.

[APPLAUSE]

Wow, great to be here. Thank you. I'd love to think that I can live up to the expectations. But I'm thrilled to be here, and thank you all for coming in this late time of the day. And I'm particularly thrilled to discuss this paper that is a work in progress. And I'd love to use this opportunity and try to go through this really quickly, so I can hear what you have to say about this.

I'd like to start with the two caveats, and that is one that is that I landed this morning at 5:00 AM. And I'm practically asleep now. This is midnight for me. And so I hope to keep being coherent.

And the second is that I am going to discuss content moderation by platforms. But I'm not going to discuss the Elizabeth Warren vs. Zuckerberg debate regarding whether platforms should engage in fact-checking and filtering fake news, over-citing some propaganda, or advertising by political campaigns of politicians.

I do have an opinion about this. I think that maybe tomorrow at this exactly the same time, we may have an opportunity to discuss this in another great event organized by the Berkman Center. Sorry, this is also a promo, not just a caveat.

But I am going to discuss another way in which platforms are actually shaping our public sphere and using power, exercising power, that they have over data and information flows by filtering, filtering the public discourse.

And this type of focus is actually related not necessarily to the, who is doing this, but to some extent, also-- but it focuses more on the how, how it is being done. And it's linked to a more general agenda related to governance by AI.

And I think that the way in which content is being filtered by a AI system worldwide, especially by these mega platforms-- this is something that is troubling, interesting, and challenging, the way in which we think about the public sphere, but also about the law and how the law can address governance by AI in general, where the law is in some ways incompatible, [? whereas ?] the way in which AI is governing behavior.

But also, it is challenging the law, in particular, when we talk about content moderation by AI. So I'm going to briefly talk about this, just say a few words about how we got here, what type of systems are out there, what type of challenges it [? raised ?] for the type of oversight and interventions that we have in law. And then come quickly to my proposal and get your feedback on this. So that's the plan.

And when you look at this slide, this is sort of the announcement by Facebook about how they do content moderation. And we talk a lot about moderators and how they suffer. And they have to look through all these materials, and that's of course valid. But the vast majority of content is being filtered by AI, automated.

And so if you can come to think about it in terms of terrorist propaganda, for instance, what is defined as terrorist propaganda? 99.5%, based on the reports of Facebook, is actually being filtered by AI before it is even being perceived by anyone, 38% of hate speech. These are enormous number. This is a robust system that is actually filtering out a lot of content.

And the question is how we got here. And to some extent, it's obvious. This is a volume of information that has to be screened and filtered. But the law was helping this development in many ways.

So one could think of this legal regime that we have adopted in the late '90s, the Safe Harbor regime, where actually immunity from liability was offered to hosting facilities if they implement a notice and takedown regime, where right-holders could issue a notice. And then platforms would have to expeditiously remove that content.

And so this has quickly turned into something automatic, first by right-holders, identifying infringing materials and sending these notices. And these robo-notices were actually dealt with, or accepted, in the accepting [? or ?] the receiving side by other robots that started to manage these large volumes.

That brings me to the current European regime, just the controversial Article 17 in Europe that is actually holding hosting facilities liable for infringing content that is distributed or shared by the users, except when they acquire a license or they install a filter. And these are strong incentives to actually implement some of these systems.

What else comes to mind is the duty to remove illegal content. In Germany, that is the Act to Improve the Enforcement of Rights and Social Networks. I hope that this this is the right translation from German.

But if a platform is being faced with a 50 million euro fine, and to the extent that it doesn't perform this removal within 24 hours of what is considered illegal, that requires some sort of automation. And of course, a removal by an hour would require some sort of automation.

And that is a proposal that passed the European Parliament, regarding terrorist content. It's still pending the approval of the European Council. So when we talk about content filtering by AI, it's everywhere. It's everywhere in copyright.

So I think content ID is the most well-known system that was developed by YouTube, started as the filtering for copyright infringing materials and then was turned into a business model of content ID, where you don't need to actually remove infringing materials but could actually monetize them if you want, depending on the choice of the right-holders.

We have a [? scrib ?] that is this book repository that is using content ID that is actually an ID, a digital ID, that is based on a semantic analysis of the book and would allow you to automatically identify infringing books. You have something similar in Flickr to identify infringing photographs. You have an AI system that is being used by Amazon for brands.

But it's not just intellectual property. So Pinterest is actually using these systems in order to identify, analyze, and remove videos of people that are harming themselves to prevent suicide. We have that in the context of removing videos of shootings, of the massacre in New Zealand, but also now more recently in Germany, a lot of reports on that.

Tech Against Terrorism is actually a consortium of all high-tech companies that developed a confidential data set of hashtags that are identifying terrorist propaganda and are using this to remove whatever is considered a terrorist propaganda by these companies, and that includes Microsoft, Facebook, and some other giants.

It doesn't have to be done by hashtags or an ID of the content that you're trying to remove and targeted by these systems. You can actually use AI in order to sometimes predict whether some content is going to be uploaded online.

So this paper is actually describing a way in which AI systems can actually analyze chat rooms and discussions among people who are planning to Livestream some recordings of sports game in order to prevent a copyright infringement. But you could think of these systems working to prevent some other live broadcast of protest, of demonstrations, and of course of violent crimes.

So we have a lot of systems of that sort. And when we come to think about this, all of these systems actually have some things in common. And so when we think of these systems, they all take content that is being uploaded by users.

They analyze this content by using a screening algorithm that is informed by features provided in the context of copyright infringement by right-holders in the context of terrorist propaganda, could be informed by governments. And these features and the way that is given to them is actually allowing the analysis of this content.

So it's either by using the hashtags or content ID, or what have you. You have some sort of outcome for such an analysis. And that is translated to some action that could be a removal of that content from the platform. It could also be the blocking of a link, if you're Google. And it could be an update to a filter that would not allow similar content to be uploaded.

What is really interesting about this system, or these systems, that makes them machine learning or AI, in the meaning of machine learning, not in the broader sense of AI, is that it has this feedback loop. So what you're removing is actually being used to inform the algorithm of what else would have to be removed in the future.

So the more content that is infringing with your removing, your algorithm will change and adapt and be refined to more accurately determine the removal, or the infringing nature, of more content that is similar. And so the more you remove whatever it is that is described as terrorist propaganda the more the system learns how to define similar content in order to remove it.

So when we think about these systems, the first function is, of course, applying a norm that is just defined somewhere. So in copyright, it's really should be easy. But, of course, those of us where I'm coming from, it's not hard to guess. So this is the background from which I'm coming.

In copyright, the question of how to apply the norm would be in the details. So you should not copy without a license. But how much do you need to copy in order to trigger infringement? Is it 3 seconds? Is it 30 seconds? Is it the whole copy?

So these norms are not only applying a norm that is already written but also interpreting the norms. And in many ways, are also setting the norms. And so that if you know that you cannot upload something that is 30%, or 3 seconds, similar to a content, you can no longer upload videos that qualify as substantially similar to the content, as such.

Substantial similarities would normally be a legal test that will be decided by courts, not by coders. And so what is really interesting is to look at this process of norm-setting, this process of norm-setting is actually defining not only what is infringing, or what is considered a terrorist propaganda rather than an expression of protest, but also has to optimize a particular goal.

So if my goal is to maximize the removal of infringing materials, that would be one goal. Maybe I can also refine that goal and say, well, yes, just infringing materials. But provided that they are not fair use or materials that are being used for educational purposes, just to make this simple, right.

But then I'll have to decide how much educational it should be or how much of a fair use it should be. And all these decisions have to be made ex-ante. You can't wait for the case to come. You have to say, how much was copied and what type of weight you're going to give to that fact that a lot has been copied, or little has been copied, compared to the question of where it came from.

Same thing with the location of terrorist propaganda, to what extent you are going to consider this when you're removing content. Once you decide what that trade off is going to be between the different values that you put into this bucket, these would be the trade-off that would be implemented by the system.

And the system optimization would implement the same trade-off. Whereas in law, we would normally think of the principle, fair use, or national security and free speech. And we will have institutions that will determine these trade-offs down the road. We won't have to do that ex-ante.

But we think that there are institutions, courts, for instance, that could decide it later on. We have these principles in the Constitution. We don't decide the trade-offs ex-ante. Definitions of the trade-offs would be concealed. We won't have any access to it.

Sometimes, even the programmers, depending on the design of the system, would not know exactly what the trade-off is, unless you are playing with the system and learn what it does. And the feedback loop would make this dynamic. So that is a system that is changing every time you get new content that could actually refine the way in which these systems are making the classifications.

This is not an error-free system. And so you can have, for instance, I think that this was hilarious when a script that I had just mentioned before is in storage, right, it's a hosting facility for books. And there was the Mueller Report that was, of course, a report that was prepared by the federal government. It's public domain. But some publishers also published it. And so they include it in the ID, right. The system actually recognizes it as something that is proprietary, and it has been removed automatically because that is how the system works. But also in more tragic situations, where you think of the way in which YouTube is actually recognizing some of the films that are being recorded-- the videos that are recorded by activists in Syria on war crimes are being deleted just because they're misidentified.

So when we think about oversight, one reason to do it is efficiency and quality control to prevent these errors and biases that are inevitable. But we want to make sure that there is a check on that. But there are other reasons why I think that we need to have oversight over the way filtering is being done. And one is that these systems that are being used-- and in this context, it's not just AI that is being done by Flickr, for instance, but by the major platforms. These systems are actually converging some of their private interests and the public function of enforcement in the same infrastructure, and with the same algorithm, with the same training data that is actually performing three different functions.

So the first functioning is the matchmaking of content and users. That is what Facebook is doing for a living, to match the users with particular content so that each of us would get to the feed that we deserve. Or that YouTube would give us the recommendation system that fit our own preferences. That's the business model. But at the same time, that same system, with the same datafication, and the same training data, and the same feedback loop, would also do-- and law enforcement, for the purpose of incitement for violence in the United States, or the purpose of hate speech if you're in Europe, right, or the purpose of copyright around the world.

In the middle, there will also be some content moderation. And that would depend on the community guidelines of each system. And so our public sphere that is made of all of that is actually the output of a system that is doing different functions. But these functions are not separated-- are done by the same algorithm data, training data, and the same learning and feedback mechanism, and could not be actually separated. It is the way these systems are working now. And so efficiency aligned with incentives, but also the need to restrain power. And here we have-- the system off of content filtering is pretty robust. And the way in which we normally would restrain the states [? or ?] separation of power, using the rule of law, with the separation of power between the different agencies that sustain our constitutional rights. That would be one way of restraining power. The other way would be a market mechanism, either in consumers rights or competition in the market.

One of the problems that we're facing, that we had [INAUDIBLE] filtering, with none of these actually being functional. And that has to do, again, with the fact that these systems are working-- are being exercised and applied by the social media platforms. So we go through the [INAUDIBLE] oversight. And we know how [? and why ?] can we have some public conversation about things that are being filtered out by agencies.

First of all, because a lot of them are filtering things before they're being uploaded. But second because the way our public sphere is designed is that when we talk about a public sphere, it's we have a conversation like we have here. But it's not really a public sphere. It's actually publics that are made of our satellite feeds. And so, we know now what we know now. I don't know whether to think the reason that I'm seeing things is because this is the way my feed-- this is how my feed looks like, or whether it is because none of us see that. So I don't really know there's less of a public-- also because we no longer have that public.

Passive view something that is talked about a lot when we talk about AI systems. It's all buried in the data and the algorithm, the living algorithm. It's dynamic, it's changing. Of course you know that what happened yesterday is not necessarily the latest situation today. Because of that, [INAUDIBLE] feedback loop.

And finally, in that public private fusion but they are sitting on their balance sheets before what we have is our intellectual property rights or just property rights over the servers. They don't want to let you in their data because the algorithms are kept [INAUDIBLE] as the trade secret, et cetera.

All right. So how do we guard the guardians? In the literature, we have a few proposals. You know, I think that it's nice to divide them into two types. Some are regulatory where you have calls for more transparency, more auditing, due process in terms of allowing more [? appeal ?] to have some way of actually dissenting a removal of some sort. I am happy to talk more about this in the Q&A, but I think that none of them provide a good solution for where we're at.

There are some technical proposals. Some are really interesting, such as requiring by regulation that platforms will reconfigure. And so we can say we know what the tradeoff should be. We want terrorist propaganda to be removed as long as it doesn't violate freedom of speech, and here is a formula. We will tell you, Facebook, what you have to do. Let's assume we're not in the United States-- and some European countries actually have an idea of how you should balance this.

But even if we had a way of telling platforms how to do that, we don't have a good way of [? oversighting ?] this. And that is where the problem is again, right? That we can tell them what to do, but we don't know how to check whether they're actually doing what we ask them to do.

Some of the subversive tools are also really interesting, a lot of proposal on how to challenge some of these systems. But they're good in terms of protests. They're also important in terms of challenging these systems, and sometimes to reveal what they're doing. But they're not good enough as an overall solution. And of course, here in the proposal [? of ?] Facebook, to have some independent oversight group-- I can talk for hours about this. Again, maybe it will come in the Q&A.

So what is my proposal? My proposal is pretty simple. And it's a proposal to actually introduce adversary into that monolithic system. And so the idea is pretty simple, is that right now you have that system, but it is monolithic in the sense that it is optimizing one value, regardless of how many values you have in the bucket. You actually decide what it is that you're optimizing, how you trade off between them, and then you decide whether to keep it or remove.

My proposal is that before you act on removal, you create an adversarial intervention by a public AI. And I'd like to talk about the public AI. I know that we are all getting used to the idea that the public cannot do creative and innovative things. But let us be reminded that the internet was developed by the public.

So the public could do a lot of things. And I think that here the public could actually develop an AI system. There many barriers to that. But one of the reasons that I think that we could do that is that if we are able to require platforms to give us the data about what they remove and run it through a system that can actually screen that decision about removal by an algorithm that is informed by the public values, I think that we can make some progress.

So first one of the questions would be what does the public value? In a very simple system, as I describe here just for the purpose of the demonstration, we can think about copyright. If we can copyright, we think about a removal system that is giving more emphasis to [? rightholders' ?] interest in view about what has to be removed. The public says something that contacts could actually include selling data and values that are not being represented here, everything that is [? externality ?] for this system.

This system could be informed by court cases about fair use. This system could be informed by observational data of libraries and schools about what is considered fair use. And we could use that data in order to teach that system. And that system-- and I think that is actually the idea-- would have to use the output of the private system as an input, make a decision from a public perspective, and then fit it into that feedback loop so we have a way that could articulate the public view in an algorithmic way.

Adversary is something that is important both for law and for computer science, it turns out. I was surprised. I come from a legal-- my background is law. And in law, especially in common law, we cannot even determine what the truth is before we have two sides. It's like judges would have to listen to the plaintiff and then to the defendant. It's very hard to even determine what is the correct and right description of the facts before you heard the other side. It turns out that in computer science, there is also literature about adversary systems that you don't really understand, but you use another system that is also wrong and you don't understand, in order to understand that first system. Because it helps you flesh out where the errors are and where the vulnerabilities are.

So adversary would help us [? oversight ?] the private system. Another issue to flesh out is data. The fact that we don't have [INAUDIBLE] data today is something that is important both for the purpose of [? oversighting ?] platforms, but also for the purpose of innovating. And here we have a way of not just reporting the data, or sharing the data with your competitors-- that is something that no platform would be willing to do-- but just running that data through the algorithm would actually allow us not only the oversight, but also the ability to innovate and build a system that can articulate public values using that training data.

The final point is about the tradeoffs. So if we think about, for instance, again, let's just think about the copyright example. You have a film that is being or a video that is being uploaded. The platform is looking for infringements. If it's not infringing, it remains online. If it is infringing, you have to run it through the public AI system. And then that public AI system may decide or may determine that this is fair use. What [? happens ?] then when you have controversy-- a conflict-- between the public and the private system?

In that case, the proposal actually seeks to resolve that conflict. But you can resolve it by a human review. But you can also resolve it in a computational way. So in some cases, we could actually think of-- I mean these systems actually don't tell you whether this is infringing or not, or whether this is fair use or not. But the output of AI systems would be it's 87% that this is infringing [? copyright. ?] Or it's 37% that this is a fair use educational purpose.

And then you can at some point create a matrix that would allow you to articulate these tradeoffs in a computational way. So the more cases that would come to a human review, you will be able to generate some tradeoffs that are pre-determined and actually can fit into the system.

But the advantage of having an adversarial system like this is actually in making the tradeoffs that the AI filtering system are making more visible. So that we can see what it is that we are missing by this monolithic system. Right now we don't know what is being removed. And we don't know what the tradeoff is. And so as an institutional structure, this system can actually enable this.

All right. So this is a proposal at the regulatory level. The idea is to incentivize platforms to run their data removal through the public AI system to allow us to build this. And here I think a good incentive would be to make the immunity or the safe harbor that they have now conditional upon running their decision on removal of that system before taking action on removing the content.

It also [? includes ?] computational dispute resolution and the human review that I've just described. And at the technical level, what we'll have to build is a public AI that would offer a real-time check on content moderation. This is a way of not doing it before the removal once or before giving [INAUDIBLE] system, allowing the system to work once [INAUDIBLE] or checking it every three months, or checking it once a year, or checking only the outcomes, or getting reports.

But having an [INAUDIBLE] system where a public system can actually check the system on an ongoing basis. Some advantages is to actually enable a more pluralistic system of filtering content where we have more values than we have now, or especially more values in the public sphere that we can actually discuss, and negotiate, and have a conversation about. Whereas now this is all being done under the cover of code. We have a public fix here for something that is being done in private. It's ongoing and dynamic. And it requires us to think more creatively about the way in which our public system and our legal system intervene in these sort of private or semiprivate public spheres.

There are, of course, some challenges. Incentives and funding-- I thought this was the biggest challenge, but now I've come to think about tax. And this is sort of a pollutant that comes from social media. So maybe they have to pay for this, but not to build this. We could fund it by using tax money on platforms in order to sponsor a public AI system that would [? oversight ?] the private filtering systems.

Of course, there are questions about what's in and what's out in the public AI, how do we determine this, who is deciding this. There are ways to do that. It's not as if we don't know how to involve stakeholders in administrative, legislative, and legal decisionmaking processes. We have that in the environment context, and we have it in other contexts where we actually have some ways of involving stakeholders.

What are the institutions and agents that are part of the decisionmaking in particular content filtering contexts? And I think what is really interesting is to think about some of the implications for law and how the law should change its role here, the legal intervention in terms of a system would require courts to actually undertake a different role. And that would be to provide some oversight through the AI public [? oversighting ?] tool.

So these are [? reports. ?] I look at the time so I really want to keep some time to hear what your thoughts about this, and so I'll stop here. And I'm looking forward to your comments.

[APPLAUSE]

[INAUDIBLE] That's fine.

Yeah.

[INAUDIBLE].

Thank you for your interesting talk. I have never thought about a public AI before. I spontaneously have two issues with it that I would like to raise, and maybe you can rebut them. The first being that right now I don't really see the additional value of having two checks, first a private and then a public check. Because it seems to me that at this point most private providers are actually not very much in favor of deleting a lot. They're basically doing it mostly due to public pressure. And so that means they're actually basically only deleting what they have to delete according to public values anyway. So why would these two algorithms actually be different?

And also even if there are written differences, why doesn't the public just provide their algorithm to the companies and say, you have to use it? Why would there have to be two algorithms?

So that's the first point. And the other point is that well, it kind of seemed to me that you were working on the assumption that there are a certain public values that we can employ in this public algorithm. But I think there is actually a lot of debate about what should actually be deleted and what should be kept on the internet. And there are people that would delete much more hate speech, and others that say we should leave the conversation much more open. So I think there would have to be very much an [? active ?] political decision on how to [? configurate ?] this public algorithm. And I don't really see any even near consensus happening about that in the near future.

Yeah. Two excellent points, thank you. So for the first point, I think, you know, you sort of assume that the platforms are removing the thing that they should remove. I would argue that we don't have a clue about what it is that they're removing. And so every once in a while, we get some anecdotes about what it is that is being removed. But except for people that are working in these companies, and in some occasions that I had a pick on what it is, I don't think that as a public, as like the polity, we know what it is that is being removed, and what are the reasons.

And I was trying to explain there are many reasons, some of them legitimate. If you're a platform and you want to maximize the number of users and you want your platform to be attractive to a big enough number, sometimes you will remove things in order to cater for the preferences. And I think we would normally think about this as a legitimate business interest. And that would depend on the country. Different countries would have different rules about the limits of free speech.

We at least have some consensus, at least [? within ?] our country, about our laws and how the law actually has to be implemented. You're right. And that goes to your second point is that in the current situation of liberal democracies is that maybe there is no such agreement, and that cannot be resolved by this system. This system can actually be good for the context in which we do have agreement. But I think that it could help us reach agreement if we knew what it is that is being removed. And I think that one of the problems that we're facing now is that this is all being done behind the [? scenes. ?] We don't know.

And I think that that creates another level of risk for really liberal democracies. The platforms can actually, if they don't have to face any public scrutiny because no one knows what it is that is being removed, they become more vulnerable to those who can know, which are governments, or you know, more powerful players in that context. So I think that just by creating a way for us to [? oversight ?] that is more practical and more visible, could be more visible for the public, I think that that is something that could also contribute to our conversation that hopefully will end up in some agreement.

But just to [? address ?] the last point to that just to make it more concrete, so in the context of copyright, we do agree. I mean, there is a law. Some people think that it should allow more fair use. In terms of terrorist propaganda, I think there is also some agreement. On child pornography, there is also some agreement. So there are some cases, even in this country, where you can agree. And I think that you will find more consensus on a more wide variety of issues in other countries outside of the United States where the issue of free speech regulation is a little bit different.

Hello. I'm [INAUDIBLE]. I'm a former member of the European Parliament, and as you know, I spent quite a lot of time trying to discourage the use of these technologies for copyright enforcement. But you are right of course. It's a fact of life that platforms do use them, and possibly have to use them to comply with the law. So I think it's interesting to think about how to make the system better. But I do see a few issues. Some are particular to copyright, and some are particular to AI.

First of all, I think in order to build such a public AI, you would have to have copyright registration. Because the example that you give, for example, of a script [? deleting ?] the Mueller report, it's a case of copy fraud, where simply a rightholder registers something that they don't actually own the copyright to. And I think that as long as you don't have an authoritative public registry of copyrighted information, an AI would not be able to learn that. Because there is simply no basis for knowing who the real rightholder is.

The other is a bit more a problem with AI as such, which is that certain distinctions are easier to make for AI. So it's easier to match a pattern to see, this song is the same as that song, even though some changes have been made. But it's much more difficult for AI to do something like determine whether something is a parody, because it's much more complex. AI would have to develop a sense of humor, so to speak. And perhaps connected to that, it's also the problem that the fair use is only a defense. So that means a platform that found it difficult to comply with fair use could simply, in its terms and conditions, say we only allow licensed materials. So I think it would also be necessary to turn the fair use of the copyright exceptions into users' rights that they can actually positively rely on against the platform.

I'm not sure that the change to Safe Harbor in the way that you propose is a good idea because, at the moment, the Safe Harbor is creating an incentive for the platform to leave things online that they might otherwise delete. And so I'm not sure if it makes sense to say, OK, you can only rely on the Safe Harbor to leave things online if you first use this public algorithms [? when ?] the process is also to leave things online because the incentives are not really going in opposite directions. I don't know. It might be a bit-- not a fully developed thought, but I don't think that the incentives are pulling in the same direction there.

And finally, I would like to question whether there is a consensus on copyright in the sense that I believe if copyright were perfectly enforced on the internet, society would collapse. And quite often in the discussion in the Parliament, certain concerns were disregarded by saying, well, but nobody is going to enforce it. For example, taking pictures of public architecture, and things like that. And so I would question whether perfect enforcement of copyright is even something that is desirable.

Well, thank you. Very, very interesting and provocative points, I think. So I'll start from the end. This system doesn't intend to improve copyright enforcement. It intends to correct it and make it more accurate since this system is going to check whether you were right in removing something. I actually think that that is something the platform would welcome.

So I mean, I don't [? see ?] any reason-- you know, if I was Facebook, why not? I think that if I hear the platforms now, they say, regulate us. Don't make me [INAUDIBLE]. I don't want to make these decisions for you. You cannot agree on your free speech boundaries or limits? We don't want that to be our problem because we have a business to run. Go sort out your issues and tell us what to do.

Now this procedure actually intends to tell [? what ?] you do because you cannot do it up front. You don't know how to do it up front. You have to engage in a conversation. And that conversation has to be computational because this is how that system is working. And in order to fix this, because now it is removing things-- I mean, you talked about the Mueller report as if it was fraud. But I didn't think it was [? fraud-- ?] maybe it was intentional. I don't think so. I just think that that is how the system works.

If you are a publisher, they assume that you are the rightholders. And that is the status quo now. If you want to fix it, then you have to intervene in it. I mean, this system cannot make things worse. It can only say it's a parody. If it doesn't recognize a parody, well, this would have been removed anyway. And so if it does manage to identify it as a parody, then it may remain. It may make things better.

So I think that the hope is that a system of that sort would allow us to articulate our public values in a way that would be effective in a computational conversation that is actually now constituting our public sphere. And in terms of context, actually, this is getting better. It's true that there were a few reports about AI that systems cannot identify context, and copyrighting doesn't identify parody maybe. But when we think about human rights, it's even worse when it gets it wrong.

But I think that it is getting better. And again, the idea is to have it informed by a variety of values and not by a monolithic set of values.

Hi. Thank you for that presentation My question is about maybe one assumption, or maybe it's not an assumption, I don't know. But you were saying that one possibility would be for the government to tell companies how to regulate content. [INAUDIBLE] oversight so [? this ?] solution would overcome that oversight problem. But it remains the idea that making moderation better means making it more similar to the legal system. And I wonder if that's-- I disagree that platforms try to take down as little as possible. I think they try to take down all the legal content and a lot of unlawful content. And for some cases, I wonder if they might have a justification to do that. According to this model, it seems to imply that a better world would be a world where Facebook allows all nudity on the site. which might be the case, or might not be the case. I was wondering if that was the assumption, and why.

Yeah. So again, I think these are very good points. And I think that there is no way around resolving some of the conflicts that we have in society about what should be, you know, part of our public sphere and public conversation. We need to agree about that. And if we cannot agree about this, no system, institution, or computational way of removing or keeping things would help us. I think that we are not moving forward in developing a conversation about this if this is happening in a way that is not accessible for us as public. We don't know, so we don't have the way of even having a conversation about this. It's not as if I think that more things should remain online or less things should remain online.

I think that the question is why, why they are being removed and why they are being kept online. I think that right now even mapping the ways in which the systems that are on the ground are being informed by different considerations, I bet that some people here also had some conversations with platforms. There were really interesting consultations that Facebook was doing around the world considering their oversight board as if, you know, that would be a solution. But you'll have a committee that would think about something that happens every-- this is happening instantly. But you'll have a committee of people that would think about what principles, how are they being implemented in the details of your system that is actually deciding whether my presentation remains and yours is being removed. This is the type of oversights that we need.

But we need that in order to decide whether more things should remain online or less things, and in order to understand why and deliberate on this. We can't do that right now.

[INAUDIBLE]

Thank you so much for the presentation. It was really insightful, at least for me. And especially the fact that it is really a challenge to just leave in the hands of any social [INAUDIBLE] intervention any kind of content moderation or the change to any automated content moderation. That's why last year, I was part of one of the dynamic coalitions of the Internet Governance Forum, and we developed a set of best practices where we were investigating how different platforms actually remove content and also how different platforms delete users, and the idea of these best practices was actually to instill the kind of due process so that users have a way of constant [INAUDIBLE] automated removal, as you rightly put it.

So of course in this case, this model that you propose seems actually quite useful and handy. However, my concerns are more in light of how to implement it. So for instance, I cannot remember your name, but--

Harold.

So as Harold just said, the decision is also very political. So even though we might agree that in certain jurisdictions-- I'm going to use the word jurisdiction as opposed to how in the internet jurisdictions do not entirely exist as such-- how to choose, not only the social values on which we can more or less agree upon, but also in the event of clashing values, how to choose which values are we prioritizing, especially if we are proposing a model that is based on machine learning. You are feeding the machine what type of values are going to be prioritized in that particular type of conflict. But probably you want to prioritize that value in a different type of conflict. So how does it work on a case by case basis is basically the practical implementation question that pops in my mind when trying to figure out how your proposal will work in practice. Otherwise [INAUDIBLE].

May I suggest, given the time, that we pick-- there were a couple more questions. And then you collect them together.

[INAUDIBLE] I need to right now?

Will that work for you?

Yup, absolutely.

I see there's one here and one right there.

Instead of one public AI, should there be three or four or more, which are written independently by people who don't know each other, and therefore have different algorithms in them?

Sir, I'm sorry. Could you just repeat this? I'm sorry that I--

Should there be more than one, three or four competing public AI systems?

Yeah, OK.

That people would choose how? Who would get to choose which one?

Well, I'm not sure. As far as speculating here, it's like, you know, it seems like it's better if there's a multiplicity of these systems and that they're written independently of each other, you know. There are other parts of computer science where we like the idea of there being redundant, separately written systems, and maybe this is another such place.

I feel like we keep upping the ante as we go around. I was going to ask how you would feel about having multiple AI systems as well. You could imagine having an AI system which is the legal floor, which is a minimum on what would need to be taken down, which still has all the oversight transparency problems that we have anyway. But why don't we just have individual AIs as well, which would then provide that [INAUDIBLE] between those things which we want our own personal experience based on our own norms to dictate, versus the legal floor?

Why don't you take these three?

OK, all right. So three good points how to implement-- again, I have an idea of how to do it. In the paper I demonstrate this in the context of copyright. Because I think actually it was chosen because this is the easiest case, and it doesn't trigger a lot of political issues. So there are some [? control ?] controversies. But I think that the idea is to encourage controversy, but just to make it more visible and actually create a space for it to happen. That is, negotiating values has to be made in a public debate. I think that as we move from speech that is not being regulated, the system is more difficult to implement. So in this country, it would be difficult to imagine a system of [? network ?] regarding hate speech. Because this is a private-- this would be determined by the different systems according to the business profile.

But in Germany, where you have a law, actually that is something that would be easier to do it. The values are actually the values that are being set by law. And the way they are being applied should be the way in which the courts have done so. And so I think that I agree with you that there will be new cases that have not been determined by law, by courts, I mean. And then the system will have to actually push the controversy into a human decision maker that could actually inform the system about cases of that sort, until that other case that hasn't been sorted out would come again.

Multiple systems, yes. Actually I think we can think of a procedure. Say we have a problem with a monolithic system. Adversary could actually flesh out some of the problems and maybe you should run this by a system that is considered the [INAUDIBLE] culture stamp as being public and not private. And I can think of also implementations. Again, I can think about this in the case of copyright, that libraries would-- actually there'll be a market for this, for fair use or limitations of some sorts. And so you can think about these filtering systems that are determining limitations and exceptions or fair use or free speech in other market situations where they're not actually filtering or [? oversighting ?] what supply platform is filtering out.

To have competition may be good. To have a procedure to say, well, you could use one of these systems that are on the market, that would be sufficient in order to [? write ?] your filtering decision on. That might work, but not a personalized one. And the reason I don't think about-- I think that the whole idea is to try and create or fix a bias, a distortion that happened to our public sphere due to these filtering systems. So I don't think it would be useful to have a person. I mean, everyone could have their personalized app, but that should be sort of your personal butler or algorithmic consumer app that would cater to your preferences. Of course, there could be a market for this. But this is not the fix for our common goal in the public sphere.

Thank you very much.

Thank you.

[APPLAUSE]