MUSIC PLAYING] KSENIA BAKINA: Welcome to the Cyber Armor, a podcast which champions voices and safety of women and girls in the digital world. This podcast is brought to you by the Center for Protecting Women Online and the Open University. I'm Dr. Ksenia Bakina, a research fellow at Streamlit and law and policy at the center. This is the final episode of this series, and today we're going to be talking about misogynoir, online hate targeting Black women.

And to address this issue, I have two wonderful guests here with me today. I have Professor Miriam Fernandez, who's a professor of responsible artificial intelligence at the Open University and a responsible AI stream lead at the Center for Protecting Women Online.

Her research agenda revolves around advancing responsible AI, ensuring that technological innovation aligns with ethical principles and societal values. Her work spans diverse domains from algorithmic transparency and fairness, to the societal implications of AI deployment.

By integrating cutting-edge AI techniques with a human centered approach, she fosters solutions that prioritize social responsibility, transparency, and inclusivity. With a portfolio of more than 100 scientific articles and having won numerous external grants supporting her research, Professor Fernandez has significantly influenced the discourse in the field of responsible AI. She is also an equality and diversity champion for the Open University.

Also, I have with me Dr. Joseph Kwarteng, a research associate at the Open University and a leading voice in responsible AI and social justice. Dr. Kwarteng recently completed his PhD in computer science, where he pioneered groundbreaking research on understanding intersectional hate speech online, particularly misogynoir. His work sits at the intersection of AI ethics, computational social science, and educational technology.

Currently, he is developing generative AI systems for smart assessment and guided education, with a special focus on reducing bias and promoting fairness in AI systems. His research has earned him recognition, including the Open University's Research Excellence Award runner up and a best paper award at the IEEE, ACM international conference.

[MUSIC PLAYING]

So to begin this conversation, I want to first ask you, what is misogynoir for-- what do we actually mean by this term? Because it's a term that many of our listeners may not have come across.

JOSEPH KWARTENG: Yeah, I mean, I would say quite fairly new term, but it is not that new. So misogynoir as a time as to describe the specific kinds of discrimination that Black women experience at the intersection of racism and misogyny. So the word combines misogyny, which is hatred against women.

And then the "noir" is the French word for Black. So the combination against that to links to hatred against Black women. So it captures the forms of hatred and stereotypes of violence that are directed specifically to Black women. In terms of what it means, I would say.

I mean, unlike general misogyny that affects all women of general kind, but misogynoir is more focused on targeting Black women and the combination of their race and gender. So it's looking at them from the facts that they have these intersectional identities as being Black and then being women as well at that point.

And it looks at how the media portrays, policing how Black women's emotions are being triggered and how these beauty standards look at all these dynamics in the online space or even in the offline space.

So we can look at stereotypes like the angry Black woman, or how people tend to see Black women as aggressive or unfeminine, looking at their body structures and features. So that sort of things is what composites, what misogynoir could be or is.

KSENIA BAKINA: And do you know when this term-- because you said it's a relatively new term, when did it develop or who actually came up with this term?

JOSEPH KWARTENG: The term was coined by Dr. Moya Bailey, who is a Black feminist scholar activist. She introduced the term somewhere around 2010, and it became more popular around 2018, where the conversations around experiences of Black men became quite popular around that time. So she coined the term when she was looking into how Black women are represented in the media, example, like pop culture or hip hop culture.

So she was looking at how the unique blend of these experiences that Black men go through and how it's been portrayed in the hip-hop culture or the digital media how Black women are shown in their bodies being oversexualized, and how people talk about their structure or their body figure and how that dehumanizing activities that happened around that time and how people portray them in the media. So that was around that.

And today, it's widely used to talk about the intersectional realities of Black women's life and their experiences. So it's that how the term came about.

KSENIA BAKINA: Thank you for explaining this, Joseph. And so I'm just wondering, do you think there's a real need for a specific term? Why do you think simply calling this, for instance, online misogyny or online hate speech, why there's a need for that specific term, identifying specific experiences?

JOSEPH KWARTENG: Yeah, I mean, thank you for this question. It's quite interesting because many times people wonder why do we have to name things. But I think without the time, the specific experiences of Black women would be lost in the broader conversations of maybe racism or sexism, but then it wouldn't be able to capture the direct or the specific experiences that they go through. Misogynoir, as a term looks into the distinct patterns that include hypersexualization, identification of Black young girls.

We can look at an example is the incident that happened around the UK about this Gen X child that was stopped and searched in school without any parental guidance, because apparently some of the teachers thought she looked a little older. And these are some of the experiences. But then this would have been triggered and put into a broader conversation of racism, which doesn't really fit in there, because this is coming from a Black young child.

By naming it, we begin to understand that something like this exists, and then we get the mandate to look into studying it and then understanding and addressing the root cause of such issues like that. So I believe, yeah, coming up with a name for it was very necessary so that we can get to the root of it and understand that something like this exists, other than that would have just been existed around other conversations of misogyny, which doesn't really capture these experiences like that.

KSENIA BAKINA: Yes, absolutely. I couldn't agree with you more. And there's been a lot of feminists who've also said over the years that this particular experience of Black women often gets overlooked because they are either treated that they should be discriminated on the basis of their sex as women or on the basis of their race because they're Black, but because of their unique experience of not just being Black, but also being women, it creates a completely different playing field for them and completely different experience.

And I'm just wondering whether you know about any studies that may have been done that identify if actually Black women experience more online hate, or if they experience it in a different way.

JOSEPH KWARTENG: So I would say yes to that because I mean, lots of studies, even including my own research, have shown that Black women experience more online violence and hate in how frequent they are and how severe they are. And in 2017, 2018, there was a study by Amnesty that looked at the different experiences of UK MPs.

And then it came out that Diane Abbott, the Black female MP, received quite one third of all the hate that was received around that study. And it showed that it wasn't just about the hate against her, but it was more about the impact on her at that point in time. Even her own team had to tell her to get off the social media platform for quite some time, for her own safety, and that tells you how severe that is.

Even for another classical example would be Leslie Jones from the reloaded Ghostbusters movie. She just starred in the movie, and then all of a sudden, went viral, going and receiving a lot of hate on that. And you could see that this was a more female, a woman movie that they casted a lot of females or women in there.

But then she was the only Black one and she was singled out and comments were looking at her body structure, even posting images of her in relation to monkeys. And it was quite too much that at some point, she had to suspend and delete the account from Twitter, live for a bit of the social media. And these are all the things that happens around those areas.

Yes, other people receive online violence, but then the severity of it is quite immersed when it comes to Black women on these platforms, because it's not just targeting them as being women, but then it's just targeting them as being a woman and also as being a Black person in society.

And you can see the tones of racism and all these sexism combined, forming something different, which is more severe, that impacts their self-esteem and sometimes their mental health as well at some point in time. And it's causing some of these silences on these platforms, where they have to leave for their own safety, which is not good for a society where we are trying to build and be diverse as we want to be.

KSENIA BAKINA: So Joseph and Miri, I know that you've done some amazing work in relation to misogynoir and how it manifests in online environments. So could you tell me a little bit more about your research and what you did?

JOSEPH KWARTENG: So the research was more focused on trying to understand this phenomenon from a different perspective, because generally, studies that have tackled this have just looked at it from a qualitative point of view, looking like maybe getting some data qualitatively or interviewing Black women on some of the experiences that they go through.

And we took a different approach. We wanted to understand this and how it manifests on online, see how existing systems for content moderation tackles this if they're able to recognize it or not, because it's coming from an intersectional point of view. So we were wondering what could be some of the impacts of these systems.

So we looked at it from two strands. First, we explored how many misogynoir is manifested or expressed online, what are some of the public responses to that and how do people with these kind of experiences find themselves when they share some of these experiences. So in there we looked at four different cases of prominent Black women in tech, who had gone through experiences of misogyny that have shared it on the platform; Twitter-- X, formerly Twitter now. From that, we gathered these experiences and some of the public responses to the conversations that they've shared.

And then we did an inductive approach of looking at the qualitative bits and the theoretical frameworks with intersectionality on which this term was built on. And then we created a lexicon, the first lexicon of misogynoir terms and phrases. And based on that, we did-- I mean, a computational linguistics approach to try and see if we can find some of these terms within the conversations that were going on. And we did that.

The second phase was now that we're able to see some of these responses and we know how the public responds to these things, how does content moderation systems or systems that are built for hate or moderating hate on these platforms also response to these things.

Are they able to capture it? Are they missing it? If they are not able to capture it, what are some of the things that they are missing and if they're able to capture it, what are some of the things they are able to capture. So with that, we looked at two popular systems that most companies use with Google perspective and its owner. They were more popular.

I mean, perspective was more popular in the industry space. So New York Times and all these big platforms used it. Sonar was more of a research based and system that came out, and many other papers have also used it to explore detecting hate as well.

So we looked into these two ways and then try to understand how this hate manifests on these social media platforms, and what are the best ways that these systems are able to tackle it, or if they are not. And then we can get deeper into why they are not able to capture it. What could be some of the reasons and what could be areas that we could explore in terms of why that is the case.

So that's the trajectory of the research, how we went about it.

MIRIAM FERNANDEZ: Yeah, I think it's also important to highlight. So one of-- because we started with, as Joseph was saying, with four use cases that we found online of prominent Black women, when they expose the experiences of misogynoir and we observe the online response. But one of the things that we really wanted to do with this research was to capture more of these experiences on how women experience misogyny online.

So actually, Joseph created this website to try to capture-- to try to encourage women to capture some of these experiences. Unfortunately, I have to say that website was not particularly successful. So we were thinking about mimicking the project of Everyday Sexism. I don't know if you have heard about this project before.

So that was a website in which at some point became very popular, and it tried to capture all these experiences of everyday sexism of women at work, at home, and their everyday lives. So we tried to do something around misogynoir online and try to create this website as a way to capture some of these responses, but I think it's also difficult for people to maybe to share these experiences as they are in a sense, sometimes traumatic or confusing.

One doesn't really understand very well what happened, and if that qualifies as a violent or hateful act. But that was also an experience for us because we thought we were going to get quite a lot of responses, and that was not the case. Joseph did an amazing work on a third phase of the project.

He hasn't explained it yet. But one of the things that we really wanted to observe is how science machines are not able really to identify this type of hate properly. Are humans able to do so? That was a big question for us. And then how this differ between different groups.

So Joseph looked at giving the same content to four different types of annotators. So we have Black men, Black women, white men, and white women annotating exactly the same type of content, all instances of online misogynoir online, messages of hateful messages that have this combination of misogyny and xenophobic or racial hate.

And it was also very eye opening for us to see how even humans are not capable, especially if you don't come from the background. And if you have not experienced this type of hate, you are not aware, so you are not capable of recognizing it sometimes.

That's why one of the things that we tend to highlight when there is a need to create content moderation systems that have minorities assessing these annotations in which these content moderation systems are going to be based because otherwise, you're going to have blind spots.

If you are not from that group, it's not that you plan to be, you just simply not aware. And we have done other research showing that when you give annotators from a majority group, for example, white male, you give definitions of what words mean that they might not know about, then their annotations start being closer and more aligned with the annotations of the minority groups.

But it's just in many cases, they just don't know what these terms mean. They don't know the nuances of certain comments. They don't have the cultural background, and they are not able to identify this type of hate.

KSENIA BAKINA: Yeah, absolutely. And I found it really interesting in your research that you went quite deep into these different nuances. And you didn't just look at all times-- all misogynoir or any misogynoir content, but you've actually grouped them into specific types and models. And I was wondering because-- and there were terms that I've never heard of before, such as tokenism-- I don't even know if I'm pronouncing it right. Tokenism.

So I was wondering if you could maybe say a little bit of how you divided the different types of misogynoir and how this was then allocated and tested.

JOSEPH KWARTENG: So I think what we wanted to do, as I mentioned at the beginning, was not just starts as OK, this is one example of hate speech and then just go about taking data and then building systems to look into how we can detect it as most people tackle hate speech. We really wanted to get to the nitty gritty of it, try and understand how this works, because I mean, I'm coming from a computer science background.

I don't really know much about the social science aspects of this work. But then this took a different toll. And then it made me understand that it's not just about building systems, it's about understanding why you want to build a systems and who the systems are going to work for.

And that's the approach that we went with this. So we looked into scholar works of issues where the people have talked about race and gender, all the charities of the things that we've looked into from the social and psychological aspects of that, and looked at works that have shared experiences of Black women and things that they've gone through.

And then we did an inductive approach with the data set that we had collected to try and see what are some of the common experiences that have been shared here, what are some of the terms that are used around these cases and what could be built out of that. So we identified a couple, I think around six or seven themes.

So the tokenism that you were talking about is part of the themes that we built on. So it's more of looking at, OK, we put somebody in representation more like, OK, a diversity hire. We need somebody here that represents these people, but then they don't really take in the ideas that a person really shares. You're just there as a token to show that, oh yeah, for example, when people are being defensive, they say, oh, I'm not racist. I have a Black friend. So just because I have somebody in my circle that is Black doesn't make me racist. But that's not how it works.

So we can look at another team that we looked at as white centering, where white people make it about themselves when we're talking about issues of race. So they tried to express it in their terms, for example, oh, the police stopped this Black man, and then he-- why didn't he just listen to what the police were saying.

But in context, maybe he was just saying, why has the police stopped me? Why do I need to provide these documents? But then they look at it from their point of view that you just have to obey the police. You don't have to do anything around that.

So a bit of white people looking at it from their point of view. We can look at some policing that described the tone, looking at the tone that the person is saying. So you're just being angry whilst you're expressing yourself, but then looking at the context into what the person is saying, but then just picking the conversation or the sentence out of toner and saying that the person is just being angry because they are Black or they are a woman in that sense.

And then racial gaslighting, describing using white-centered narratives to downplay the validity of what is being said. And then we also looked at other opposing terms of allyship and resistance, where it was quite interesting to see that they also had a counter narratives in how people can resist these things and when they support Black women in that case.

So those are some of the things that we looked at; hypersexualization, body shaming and textualism, looking at creating some sarcasm around the bodies of Black women to make them look like, OK, yeah, I'm not really shaming you, but I'm just creating an impression of that. But these are all examples of models of misogynoir that can infer and that can affect Black women as well.

So that's how that themes and-- themes came about when we were looking into them.

KSENIA BAKINA: And I know you've mentioned that you used the detection-- the most popular detection tools. And could you just say it? Remind me what these were and how did you

actually-- did you purchase access to them or how did you actually test these terms and these types of misogynoir against them?

JOSEPH KWARTENG: So with the data that we had from the four public cases that we had, we built, I would say, potentially the first misogynoir data sets around that. So we had that data and then we wanted to understand, OK, if this is how the public is responding to some of these hates, what could be the issue with these popular ones as well? So we did a bit of a background study in looking at the scholar works and seeing what has been used most popularly in most of the academic research and also in the industry as well.

And then we came across Google Perspective, and then HateSonar. We also wanted something that was more open source that because I mean, when it comes to some of these works, you need something that is open source that people have access to and that's popularly been used in most of the industry in the cases, so that you could bee-- give a generalized form of results when you're looking at that. So with Google Perspective, we were quite lucky. I reached out to the team at Google, and then we let them know that we're using it for research purposes. So they were able to give us access to that.

And then we've HateSonar as more of a research open-source tool. So it was quite open. So we also had access to that. So what we were trying to do is these had different models. So Perspective was looking more into toxicity and looking at the definition of what could be toxicity. There's an element of misogynoir in there because they're looking at violence against a group of people and how they had defined it.

So we gave it our data sets that we have annotated as potential cases of misogynoir, and then we asked it to OK, how-- in what range-- because they start from 0 to 100-- in what range would you rate any of these tests, in the cases of it being a potential case of misogynoir or not? So in that sense, whether it's toxic or not.

And then we asked the same thing to HateSonar as well to rate it as well. And then we passed it to the experiment to see how much it would look into. And then the result was quite interesting, because we realized that most of them were more focused on cases where there's an explicit mention of saying key words.

So when there's an N word or a B word, or even when they are making references. So if I'm saying I'm sharing an experience that somebody called me the N word at this point in time, it would rate that as being something toxic or potential case of misogynoir. And we realized that then that could be silenced or potentially silenced in Black women on these platforms.

Because if I have this experience and I come on Twitter or Facebook or whatever the platform would be, and I say that I had this horrible experience today, somebody called me this word or that, these systems don't understand the context leading to why these key words are there. And they would just pick that phrase and this key word and then rate it higher or a potential case that it being toxic or even violent.

And in terms of moderation, these systems can just ban your account or just suspend your account in that case. So those were some of the results that we were seeing. So with misogynoir being very subtle, and most of the times it's not even about overtly having some keywords in there. Somebody can just share stuff with being the Black woman or stop being angry as a Black woman.

In that sense, these things might not be able to pick them up because it doesn't understand what the angry Black woman stereotype is coming from. It's just looking at it as a conversation or as a sentence. And that's the issues that we had with these popular systems that are used for content moderation.

KSENIA BAKINA: That's really interesting, actually, because you wouldn't expect that for these detection tools to not just-- if they didn't-- you might expect them to not pick up everything, but to them to actually silence those who were trying to share their stories and their experience of racism, that has a whole other ball game that you're not only not picking up the abuse, but you're also actually silencing those who are trying to complain of it.

And I know you've both Miri and you, Joseph, have started to talk about the results and how they compared. But I wanted to ask in terms of were there any information about which did-- which types of misogynoir were easier to pick up for the detection tools, like whether it was tone policing or tokenism? Was there any significant difference between the detection tools that you used in regards to the types?

JOSEPH KWARTENG: Yeah, there was. I think, to get to that, the issue with what you just said was most because some of these systems were not trained on a data set that was diverse, to be fair. So they didn't have any form of these experiences of these Black women in there. Where the data collected, mostly they just collect data around keywords or things that might be hateful.

It might not capture all the experiences or the diverse experiences of some of the things that Black women go through as well. So it could be an issue with the data, or it could be an issue with the annotation, or it could be an issue with how the model just learned to be able to detect these things.

And then leading up to your question, yes, we do find--

MIRIAM FERNANDEZ: Joseph, if I may interact there. It's important to think that no algorithm is perfect, OK?

JOSEPH KWARTENG: Exactly.

MIRIAM FERNANDEZ: And in the same way that we experience biases, algorithmic biases in other contexts, we do also experience algorithmic bias when it comes to content moderation. And we have seen that, there is quite a lot of research about that. And we in our group have also been doing specific research about that. And we seen that content written by specific groups, so for example, African-American English tends to be much more identified as hateful than other types of language.

So in the same way that we need to ensure that we have bias identification and bias mitigation methods in health and in finance and in other contexts, the same happens in content moderation. So we really need to ensure that. On the one hand, we make people safe. But we also don't over control for certain types of language. And these are both important aspects because sometimes people is like, OK, well, then just overremove. But it's not that simple.

So because if I overremove to avoid a problem, and obviously the algorithm is never 100% accurate, and language is very subtle, and new terms keep emerging every day because in these online communities, people have their own conceptualizations and their own realities, and they create new terms to express those realities, so if the algorithm is not perfect, and I go towards

an approach on overdeleting potential instances of hateful content, then if I am not careful, I can be biased and then I can end up deleting content of particular communities.

So in the content moderation research field, there is a specific area that is called target identification that tries to look into which community is the target of the hate and try to observe whether you are overtargeting certain communities, which is also a really important aspect.

KSENIA BAKINA: Yes, thank you very much. And actually this, as you were talking, there's a question that popped in my head that it's quite a difficult task to make sure that on one hand, you moderate the hate speech and you're sensitive to the specific context and to the subtle misogynoir hate speech, but at the same time, you don't over delete. And then I'm worried, well, being sensitive to misogynoir, will it negatively affect potentially other ethnic minorities or other types of intersectionalities that might then get overlooked.

And I was just wondering. This is my big question is like, how do we address this? How do we A, ensure that there is that right balance to begin with, but also that we don't then in focusing on tackling misogynoir content, don't exclude content that might be discriminatory towards LGBTQ+ individuals or other ethnic minorities? How do you think that could actually work?

MIRIAM FERNANDEZ: Various points I want to make. One is that the fact that you try to protect a group, that doesn't mean that you are under protecting another group. Again, that's why within the area of content moderation, there are different subfields. There is the people that study whether messages are toxic or not, and there is the people that tries to identify who is the target of the hate. And if you have, as Joseph was saying before, the way these machines work is that they learn by examples.

So what you really need to make sure is that you have as many examples as possible from a diverse group of people, diverse types of hate and so on and so forth. Because if you only give machines certain type of hate, obviously, they will become very good at identifying that type of hate and not others. So if you don't want to have blind spots, ideally you will give them lots of examples targeting different communities, different type of language, different type of vocabularies.

That's one of the things that makes this very, very difficult because hate evolves, language evolves. So it's not like a one off. I just give you lots of examples. I train you is OK, I need to keep training you and I need to keep-- you need to keep learning. It's like a child. You need to keep learning.

The other thing that I also want to point out, that is very, very difficult in content moderation is that toxicity or understanding of toxicity is sometimes cultural. And we have done this experiment. For example, this has been with radical content. So at some point, we were also studying radical groups, including the manosphere, which is extremist groups that are very misogynistic and violent against women.

And then one of the things-- so we gave exact same content to annotators in different regions of the world; so in North America, South America, Europe, Asia, Africa. The same content in North America was considered radical. Most of the content, most of the posts that we gave annotators were considered radical. In Africa, most of those posts were considered non-radical.

It's the same content. But cultures are different and interpretations are different. And when you are thinking about content moderation at a global level is very difficult because who decides what's toxic, who decides what's hateful. And the problem is that we don't really have good

definitions or good regulations to say what is hateful or not, or what is toxic or not. And at this point, the ones who are deciding these are the companies who are developing these systems.

So it is a difficult problem. And we right now don't have very good mechanisms to decide what stays in and what goes out. There is also a big issue in content moderation. So obviously, there are millions of messages that comes towards this platform every day, not millions, probably billions.

And most of that content is deleted in an automated way. So we try to do automated systems that can identify that content. But then obviously, so the things that are clear taken out, but then there are the borderline things that are unclear, and those things are given to human annotators to try to identify, whether that content is hateful or violent or not.

But we also need to consider that these are humans, and this has a big toll in the mental health on these humans. And there is not always the safeguards to protect these people that are behind content moderation systems. And I myself have been looking into radicalization and Jihadist, far-right extremist misogyny. And I can tell you that at some point, I decided not to look into videos anymore and just look into text, because it was really emotionally very draining. It was really impacting my sleep, my mental health.

So we really need to make sure that when we try to generate-- I know automated systems are not perfect, but we really need to make sure that as-- that's why we do the research we do. We need to make sure that we do them as good as we can to make sure that obviously, there are human moderators to decide on the borderline cases. But we also need to make sure that these human moderators have the safeguards in place to protect themselves and their mental health. And we really need maybe better regulations around people working in this space because it is unsafe.

KSENIA BAKINA: So Miri talked about the difficulty of on one hand, needing human moderators to assess decisions made by detection tools, but on the other hand, the negative psychological impact the moderating job actually has on them as human beings because of how much hate and abuse they are witnessing online.

So I want to come to you, Joseph, and ask you, well then, what are your thoughts on actually matters decision to get rid of all human moderators and make all of their detection tools entirely automated. What impact do you think that will have on the ability to tackle misogynoir online?

JOSEPH KWARTENG: When I saw the news, I felt like-- I think it's more of them trying to cut costs and political pressures done move to more safer platform governance, to be fair. Because I mean, as Miri said, we can't trust these systems 100% enough to be able to do the jobs that we have asked them to do.

First off, we don't know anything. I mean, it's more like a Black box in what Meta has created for their content moderation systems because we don't know what is tackling. We don't know what data it was trained on. We don't even know how these systems are built, so there is no transparency in that.

And a clear example is even when you look at X or formerly Twitter, you see sometimes you get all these things when you go through your comments or your replies, you see that this puts some things in privacy view. And then they say these things are offensive. I mean, I would say 9 out of the 10 that I've seen those things, you open the reply and you see that it has nothing that is offensive about them.

It might be that the person just wrote something random that is out of context to the conversation thread that's going on, or he just wrote something gibberish, I would say that wouldn't make sense to the reader, but it is being classified as being offensive.

And some of-- these are some of the things that why we need human moderators at the end of the day to be able to check these things. These machines, as Miri said, don't understand the culture or the history behind all these things. They don't understand the context in how these things operate. They are just given a bunch of tweets or a bunch of online picked-up messages and saying that has been annotated by a group of people that says, this is hateful, this is not hateful. We don't know any information about this group of people.

I mean, tons of research have shown, even in my own research, that I've done, has shown that person's personal belief or their identity or their perspective always influences these annotations. I may not have experienced this kind of hate, but then the paper says that anything that says the N word or the B word is hateful.

What of something that says the N and the B word, but not explicitly, but in a different format? I would say, oh, that's not hateful. And the systems are going to learn all these patterns and say, OK, anything that says the N and the B word is something that could be hateful, anything that doesn't mention that might not be hateful.

So at the end of the day, we need these human moderators to be able there to look at these other-- a pair of eyes to scrutinize what the decisions these things are making. But then if that is not there, then I think Meta does not truly care about a safer and inclusive environment. They are more of trying to cut costs and expanding their reach in profits than looking at intersectional abuse that would benefit society at the end of the day.

KSENIA BAKINA: Absolutely. I think you make a really valid point, because not only we don't understand how the systems were trained and what terms they must pick up in the first place, then you're getting rid of all of the human annotators who may have at least picked up something.

So instead of actually adding more human annotators that might be more intersectional, that might have specific contexts and specific understanding of misogynoir or other types of online hate speech, you're removing that aspect completely, and it seems that online has hate speech on the whole, misogynoir or targeting other ethnic minorities or targeting LGBTQ+ individuals is just going to go through the roof because you're getting rid of any safety net, irrespective of how bad that safety net was.

There's none at all at the moment, which is really concerning for, I suppose, future expression online and society on the whole.

JOSEPH KWARTENG: On the whole. Exactly, and I think one of the key things is I think Miri said it's up to you, but the definition of hate and all of these things we are not yet fully have a normalized time for these things. And then you go on these platforms policy and hate content page and they say, OK, we are tackling hate. And I mean, once we don't have one size fits all systems doesn't really work as we've noticed.

So if you're saying you're tackling hate, what hate are you tackling? Because there are different forms of hate. Are you looking at hate intersectionality or are you looking at hate as a single component or are you just looking at terms that could be hateful or hate targeting group of people or marginalized group or what hate?

We don't have definite definitions on their platform. They just only say that OK, we are tackling hate and then we've built content moderation systems that are going to be doing that. What are they built on? These are some of the key things that we don't have prior knowledge into them.

So sometimes we might not understand them. You might just randomly get your account moderated for spreading hate. You don't even know what you said. But then, yeah, it's one of those things that happened. And then the extra pair of eyes that could look into these things and give us a proper view of what could be taken away as well. So I don't know how the future is looking with that.

KSENIA BAKINA: In today's podcast, we discussed the misogynoir, online hate experienced by Black women. We explored the impact this has on victims, as well as effectiveness of detection tools that are used by platforms to detect online harms. And finally, we've addressed what improvements are needed to ensure that misogynoir content is tackled and addressed effectively. Thank you, Joseph and Miri for being here with me today and enlightening me on these issues.

MIRIAM FERNANDEZ: Thank you, Ksenia. It has been a pleasure. Thanks a lot for inviting us.

JOSEPH KWARTENG: Same here. Thank you very much. I think it's been an interesting conversation. So yeah. Great to do this. Yeah.

KSENIA BAKINA: This has also been the final episode of this series of the Cyber Armor, which is the podcast brought to you by the Center for Protecting Women Online and the Open University. You can stay in touch by following the Center for Protecting Women Online on our LinkedIn page.

If you want to hear more about the center's news and our research going forward. I hope you've been enjoying this podcast series. And I hope that you get in touch with us through our LinkedIn page in the future.

[MUSIC PLAYING]