Priyank Mathur

Founder and CEO, Mythos Labs

Excellencies, distinguished guests, and fellow panelists, it is my honor to address you all here today. I'd like to share a presentation if that's possible.

While that comes up just to tell you a little bit about Mythos Labs. We are an international organization that uses technology and media to combat violent extremism, myths, disinformation, and online harm.

While the threat of terrorism and terrorist narratives is nothing new, what is new is the digital battlefield that we're fighting this threat on. Today, there are almost 7 billion smartphones on the planet, and an overwhelming majority, almost everyone who has a smartphone has at least one social media app on their phones. As you can see, the average time that we spend on our phones across the world is about six hours of screen time a day, and the scary thing is every time I do this presentation, I have to update that stat every month because it's only going up.

We have noticed in our work countering terrorist narratives around the world a few key trends, a few key tactics that terrorist groups across ideological spectrum, across geography, continuously use to spread their narratives online – and those are spoofing, scapegoating, and decontextualizing. What I mean by spoofing, of course, is creating websites or online assets that are meant to mimic news websites. Decontextualizing – we're seeing all over the world where you take pictures, crop them out, videos that occurred many months ago being broadcast as though they were happening today, and other examples of social media posts being shared without any context to dramatically alter their meaning. And of course, scapegoating, which is blaming one group squarely for problems in a region, has been a very powerful tactic that terrorist groups have used across social media and traditional media as well, as mentioned by some of my co panelists.

I'd like to draw your attention, however, to two new technologies that could dramatically, we think, increase the reach and the efficacy of how terrorists are creating narratives and how they're spreading those narratives. And those two new technologies are generative AI, and the decentralized Metaverse. Just to level set, what I mean by generative AI is of course any system that uses machine learning to produce text, video images and any other kind of content. We're all probably familiar with some of the bigger models like ChatGPT and Meta's LLaMA, but there are dozens of models being made every month that are gaining in popularity that can create not just text but also videos, audio, images, all kinds of content.

The effect of generative AI on extremist propaganda that we're likely to see is an increase in the quantity of propaganda the terrorists can create, an increase in how many people they're able to reach – because you can use generative AI systems to create propaganda in any language, format, localized for any audience – we're likely to see extremists be able to create propaganda much faster using generative AI tools. And of course, most of these tools are free or low cost. So, the affordability of creating extremist propaganda is also likely to be increasing. All of that will result in a difficulty, an increased difficulty, in our ability to moderate content online.

And that's not to say that content moderation is not important. Of course, it's important, but it cannot be what we rely on alone to prevent the spread of terrorist narratives.

So what are some of the risks posed by generative AI specifically when it comes to extremist propaganda? We've seen three categories of risks emerging. The first is of course, AI generated propaganda – the use of and the creation of unreliable AI generated news sites, social media posts – but there's also fake images, voice cloning scams have picked up in nature and extremists are looking to exploit that as well and of course, deep fake videos, which can cause mass confusion related to ongoing events. The second category of risks that I'd like to talk to you about with generative AI is AI enabled radicalization. I've put up the word Eliza effect because we're already seeing that AI is going to be able to reinforce and encourage extremist ideology and behavior, particularly in audiences that are on the path to radicalization. Recently, a young man was arrested for trying to assassinate Queen Elizabeth in December of 2021. Armed with a crossbow, he entered Windsor Castle. When investigators were going through his phone. They found that he had downloaded an AI chatbot and he called it his AI girlfriend. It was a virtual companion that he would talk to every day and share intimate details with, and when they were going through the logs of what he said to this chatbot they found that he had actually told the Chatbot a few days before the attack that he's thinking of assassinating a member of the British Royal Family, to which the Chatbot replied “That's very wise, I believe in you.” So, you can see that AI could definitely, AI chatbots, that are designed to reinforce and validate their users feelings, could have a very dangerous effect when those users are predisposed to radical notions. And the third is of course LLM poisoning, which is the idea of poisoning the well. Researchers at Stanford recently demonstrated that with just $600 and 48 hours, they could create something called poison GPT, which is a disinfo supply chain attack, where you change the weight of a large language model. So it's not about creating fake content using generative AI. It's about changing the very definitions that the generative AI model is built on to make it more biased towards a certain group or more prejudiced in favor of a certain group. So these are all threats that terrorists could indeed exploit and we're already seeing in some cases, them having an effect on extremist propaganda. But what about the metaverse as well? There's a lot of talk about the metaverse – people say “Well, it hasn't really taken off yet. I don't see people with virtual reality

headphones everywhere." We forget the metaverse doesn't have to be three dimensional. There are 2D meta verses that are already extremely popular. If any of you have teenagers in your life, you're probably familiar with some of the logos up there like Roblox or Fortnite. These are very popular games that are Metaverse applications.

And of course, Metaverse societies are any virtual society that can be used to work, to play, to just hang out and are growing in popularity, especially after the pandemic with young people who are confined to their homes.

In the Metaverse, the dangerous thing, with regards to extremist propaganda, is you don't just observe propaganda, you're going to be able to experience it. So, the internet is largely audio visual, you can either see things or hear things on social media. But in the Metaverse, it'll be multisensory, you'll be able to see, hear, touch – Google is already working on haptic sensors that you can put on your body when you're in a virtual environment to feel cold or hot. Or pain, or smell things.

With the internet, extremists have to rely on storytelling to convey their narratives. But in the metaverse, they'll be able to actually create worlds. Imagine, we've all in this room probably encountered somebody who has been radicalized or greatly accelerated on the path to radicalization simply by watching a few YouTube videos or reading Facebook posts. Imagine if that person could now be immersed in an actual three dimensional world created by terrorists, how much more powerful the radicalization effect would be. And of course, the internet and social media today are largely centralized platforms. There is a person, an entity, a CEO, a trust and safety team that you can go to that you can hold accountable. But the Metaverse and the decentralized Metaverse especially, by definition, will not be centralized. It will not be run by entities and governments – there will be many Metaverses that will be run by ideologies and groups.

And of course, these are some examples. Recently in Russia, a teenager was arrested for preparing to blow up the headquarters of the FSB building in Moscow, and the way that he was practicing for this attack was by creating a simulation in the Metaverse. On the bottom I've got an example of something a user in Minecraft created, which is a Metaverse game, which is a Nazi concentration camp where people could spend time, and you can imagine the kind of effect that would have on folks seeking to spread terrorist ideology.

So how do we fight back? Three things that Mythos Labs has found effective in terms of what we've been doing: fighting viral lies with viral truths being the first. We've been creating edutainment driven counter messaging. The idea being if we are going to counter terrorist narratives, we have to do that in an entertaining way. We have to partner with social influencers with comedians, because these days the opportunity cost for watching something boring is higher

than it's ever been in human history. Why would I, if I get boring now during this presentation, all of you will be on your phones. We all have so many other options for entertainment, that if we want to captivate the attention of young people especially, we have to make our content entertaining. The second is, of course, media literacy. I think including AI literacy and Metaverse literacy is going to be key to the success of media literacy programs that prevent terrorist narratives. This is an example of a young woman I met in Indonesia, who got radicalized by an ISIS terrorist in 2016, went to Iraq, had children with him, eventually escaped and came back. When I asked her why she did it, she said, “Well, he lured me through these Facebook messages. At first I didn't want to engage.” I said, “Why did you engage?” She said, “Well, he was draining my phone battery. I had to respond.” I said, “Why didn't you just block him?" She said, “What's blocking?” She didn't know you could block a user on Facebook, if she had, maybe she literally never would have joined ISIS. So, media literacy can and does prevent radicalization.

In India we've created something called SMILE that we're proud of, the social media and internet literacy module. It is an edutainment based approach. It's a 45 minute free e-module anyone can take at SMILEmodule.com. We partnered with a very popular local comedian Saloni Guar. She explains key concepts related to extremist propaganda, mis/dis information, online harassment, and after each video, you're asked to answer a few multiple choice questions, at the end you get a certificate. And we found that people who took SMILE, it's had a huge impact in their ability to identify and report extremist content and disinformation online. So, huge promise in terms of media literacy as a way to prevent terrorist narratives from spreading. And lastly, and it's been incorporated, of course, into multiple schools across the country endorsed by UNESCO. Lastly, I'd like to talk about how we could just quickly use AI as well, not just as a risk but there's tremendous opportunities. AI could be used to counter extremist narratives. We did an experiment recently in Southeast Asia where we brought together high risk users, people who had been searching for pro-extremism terms on social media, and we showed them two sets of P/CVE content – one set generated by human practitioners and one set generated by Aldus, which is an AI tool that we have developed at Mythos Labs, which is the world's first P/CVE based generative AI platform. And what we found was that the P/CVE messaging created by the AI, which was an expert in all things P/CVE, was actually more effective, it was better received by the high risk users than P/CVE messaging created by human practitioners. So, tremendous promise is there for AI as a tool to counter terrorism, not just as a threat that we need to be afraid of. And with that, I'd like to thank you very much for your time.