Eye on Security

An Inside Look at How Reddit Fights Cyber Threats

Transcript

Luke McNamara:

Ladies and gentlemen, welcome to another episode of the Eye on Security podcast. I'm your host, Luke McNamara. Joining me today is someone that I've had the pleasure of working with and learning from in the past, Aylea Baldwin, Threat Intelligence Lead at Reddit. Aylea, great to have you here.

Aylea Baldwin:

It's great to be here. Thanks for inviting me.

Luke McNamara:

So there's a lot we're going to get into here, but I thought where we could start, because I think this will inform some of our discussion and where we're going to take it in this conversation is maybe just giving folks a little bit of a background on your background and your expertise.

Aylea Baldwin:

All right. So I have been in intelligence work in some form for 20 years, I'm a little ashamed to say that I am old enough to have been working in this field for 20 years. I started in more "traditional intelligence," I was doing signals intelligence in the military. I did that for six years before I started to kind of move into the realm of technology, where I worked for an IT solutions company for a couple of years outside of intel. And then following that when I really realized that my passion, my calling, really was intel work, I got back into intel, but in the cyber realm at the State Department, the United States State Department. And so that's where I started kind of fusing my understanding and knowledge of intel work with the idea of how this manifests in the realm of computers and networks, and really learned a lot there about where I ended up focusing a lot of my career in state sponsored activity, looking at cyber espionage and things of that nature.

Luke McNamara:

In your current focus at Reddit, describe for us a little bit about what your team's mission is and what you do.

Aylea Baldwin:

The team that I'm on in Reddit is called threat detection and we're inside of a larger safety organization at Reddit. And the mandate for safety really is something that we hold very dear and it's to protect good

users from bad experiences. In everything that we do we try to make sure that we're upholding that mandate. Specifically in threat detection, we are responsible for looking at malicious or suspicious coordinated activity, largely by external agents that are working to manipulate content and people. So really, instead of looking at the one-to-one things that might be going on the platform, we're trying to figure out the story behind more large-scale coordinated behaviors and how do we detect and mitigate those things?

Luke McNamara:

I think one of the things that's always fascinating to me when you're getting to the topic of securing a social media platform or the content and user experiences on social media platforms is oftentimes I think you hear people talk about those in a sort of monolithic way. But of course they're all very different and the sort of way that everything is set up, the different aspects of how closed off it is or the ways that people engage on the platform, I think, make securing a platform like Reddit different than some of the other social media platforms. So what are some of those things specific to Reddit and hopefully people listening, if they've not already, will go check out Reddit and all the different companies that exist there. But what are some of the unique things about Reddit and how things are structured and the ways in which people engage in conversation there?

Aylea Baldwin:

One of the things that I think is very interesting about the way that we have to balance security and safety and Reddit is understanding the thresholds and the level of tolerance that different communities have for different things. So site-wide rules can be more strict or a little more tolerant of things than others. So I think that one of the things that we have to pay attention to is that tailored experience. When we're looking at things that are reported as violations, we have to understand that there's a difference between what we may see at the platform level as a violation and what a community understands as tolerable. The other thing that I think is interesting in terms of like presenting a unique challenge to safety at Reddit is the fact that Reddit very much respects the privacy of our users. We are not the platform that really collects a lot of information on the user base and actually even refer to a synonymous where you don't even use real names.

We don't require a lot in terms of data from you to register. So as you can imagine, when doing the job of security, we have to do that with less data. And I think one of the great things is that we have a lot of brilliant minds in the problem, which really have made us able to do more with less. We're able to respect user privacy. We're able to not have maybe as much data as others, but we're still able to strike that balance and still do what we need to do to protect the experience for the users without having them kind of compromise their idea of privacy.

Luke McNamara:

It's an interesting aspect, I guess looking from the outside where so much of the activity obviously that takes place on Reddit is on the open. I think you have the ability to communicate directly with other users within the platform, right? But most of it is kind of out there being posted within these different communities within the various subreddits. And so I guess I was thinking as you were talking about that, you have an interesting dynamic where you have some of these subcultures that spring up where they have their own terminology and memes and the way that they go about kind of conversing and

engaging on the platform. But that's also part of a much larger ecosystem that kind of has its own culture too. When you're thinking about it from an adversarial standpoint of how various threat actors or malign influence could seep into that, what are some of the things that you have to think through of how the platform can be exploited or targeted by these sort of outside influences?

Aylea Baldwin:

There are a number of things that we consider. I think some of the more obvious are situational, what is going on in the world, because it is a social platform. So what are the things that could be catalysts to even make these foreign actors desire to use Reddit for influence operations? And then what communities might be susceptible to that target? What types of communities might be prime vectors for these types of operations based on whatever the social situation is at the time? We also have to think about what that means in terms of how our communities are built around the democratic voting process, for how posts are amplified. And one of the things I do want to note is amplification, in my opinion, is a lot more difficult to do on Reddit because of that voting practice. So you can't just post something and then have a billion people repost or share it with friends.

It's very insulated in a community, except for the fact that voting is a way that people can elevate something or elevate a topic or elevate a post. So what the foreign actors really have to try to do is get buy-in for people to actually upvote these posts. So I do think it adds an extra obstacle, but that works in our favor. I think that at the end of the day, it's still something we have to consider and look at. Vote manipulation is the thing that we absolutely look at in terms of understanding suspicious and malicious activity on the platform.

And I think that our structure with communities and moderators really, really does help act as a force multiplier when it comes to defense. Because these are people who have an investment in the communities, but they also understand them very well culturally and they can see anomalies much faster than we could in the sense that they know when something isn't fitting with what's normal in the engagement and interaction that happens in their communities. So yeah, a lot of times the partnership that we have with the moderators and partnership that we have with users of Reddit and understanding that they are able to directly report things, does also help us in being able to surface understand and then of course in the end mitigate a variety of kinds of threats to users.

Luke McNamara:

So we've talked a little bit about the sort of problems that you're focused on dealing with extend beyond just disinformation. But I do want to focus on that a little bit because that's one that I think increasingly over the last couple of years, you're seeing more and more people talk about and put into the category of different cyber threats that we have to think about and contend with and have their own various impacts. How do you think about some of these threats, particularly coming from your background and with your perspective focused on foreign adversaries, more in the cyber espionage space, other components of that as well, going back in your time working in signals intelligence. How do you think about the problem of disinformation and influence operations across social media?

Aylea Baldwin:

It's really something that's I think evolving even for me. I think I spent a lot of time working in much

more concrete areas of this. When we think about foreign actors and their malicious cyber activity. IO, information operations, is a whole new world that is evolving itself. And I think that having years of experience understanding factors that serve as catalysts for this type of activity, just in general, really helps to frame situations in a way that allows me personally to kind of hypothesize about what could be. So could this thing that's happening, let's say for example with the rollout of vaccines, play out as different narratives for information operations on our platform, what could that look like? What communities could be impacted? Where do we see these things that are happening in the world turn into the possibility for information operations? And then how do we think about detection for that?

How do we put in place things that can help us surface these things as they're happening? Because I think a lot of the other pressure that social media has that's a little different maybe than other security operations, is that it's an ever-changing landscape in the way that it's really driven by temporal things. A lot of times it's driven by things that are happening geo-politically, or things that are happening socially or everything from an election to COVID to something as simple as grassroots movements for a particular agenda in a certain country. There are a number of reasons why things can play out as an information operations, what's the word I'm looking for? And so I think having the back history of really thinking about the adversary, thinking about their motivations, understanding why they may do what they do, really helps, I think, in this space to kind of like I said, frame situations in those terms and try to then think through the possible scenarios and try to get ahead and kind of tease out and hunt for these types of operations on the platform.

Luke McNamara:

Yeah. Which I guess in some ways is no different or not too dissimilar from conceptualizing what spearfishing lures are going to be used in the coming weeks, if there's a G20 conference coming up or there's some other event happening that could drive that sort of activity. But one thing I want to get your perspective on, and I'm sure this has been a conversation we've had in the past before, but sophistication. So thinking about sophistication and maybe more traditional intrusion operations, cyber espionage, you can see some threat actors that are incredibly sophisticated with how they carry out the actual operation itself. Maybe the malware development, you can kind of categorize a level of sophistication around that too, how well it was developed, maybe the overall stealth of the operation. But sometimes there's a divergence between threat actors that maybe are not incredibly sophisticated with their tools, with their capabilities, how they go about their operation, but from an impact standpoint, they may still be incredibly effective. So how do you think about that concept, I guess, of sophistication and how it applies here in the disinformation space?

Aylea Baldwin:

I think that it would also largely depend on the target, which one should be prioritized. Because I do agree that there is a kind of tactical sophistication and an operational sophistication to consider. And the components for those things are very different. And one of them may be more focused on the tools and the technical ability to do something, to be able to bypass firewalls or to be able to have a very complicated malware. And then there's the other side that's really more operational sophistication that's really focused on being able to have social engineering that's unparalleled or to have operational security in the operation that doesn't allow defenders to be able to kind of trace your steps or understand your methodology. So I think that depending on what the target is, the weight of those types of sophistication will be different. If somebody is trying to get into your organization and exfiltrate

information, obviously the technical sophistication is going to be pretty heavily weighted because I'm assuming that in an organization with a security apparatus, you have things in place.

The actor's going to have to work hard to technically bypass what you've got. Now, I say that knowing that the social engineering is also a key piece of that a lot of times, with spearfishing in particular. But with information operations, I think it is more of a social engineering game. So that's really what it is. There isn't a lot of technical sophistication involved, if any. A lot of it is really understanding your target. And that usually is people. So understanding how to manipulate people and how to shift conversation, let's say in the direction you want it to go or to be able to promote a narrative that you want to promote to be able to get more buy-in is very much tied to a very high level of operational sophistication. And I think that's where in cyber threat analysis, it's a place that I don't think we focused a ton of energy yet.

I think getting people kind of specialized in the psychology of these types of operations, I think defense in this space is really tough because of that. I think that a lot of times in CTI, we focus on understanding technical operations, how networks work, how malware works. But really IO, like I said before, is really an evolving space. And I think we're starting to understand now that that operational sophistication, that really is the biggest part of this that we don't necessarily specialize in in this field is one that we've got to definitely dedicate more time and training there.

Luke McNamara:

Yeah. And one of the reasons I guess I asked that question is getting to this question around assessing the impact of disinformation and influence operations. And that's one where I think I wish kind of holistically across the space, everyone that's kind of involved or has a piece in addressing and dealing with these sorts of campaigns, we understand they're malicious. We understand that they have to be analyzed and disrupted and kind of awareness to users, to citizens, is important to make them aware that these threat actors are carrying out these activities.

But one of the areas where I think that it gets very difficult to differentiate and we sort of differentiate, or we talked about differentiating based on sophistication of methods, but how do you think about differentiation based on impact? Because I can imagine there's different mechanisms you could have to look at engagement on a post. You could have mechanism to assess how many people saw a post or something like that. But being able to actually assess what's been the impact to the individuals, to the group of individuals, that may have been impacted by these operations? That seems to be a much more difficult question to answer. How do you think about that? I'm giving you the easy questions obviously.

Aylea Baldwin:

Yeah. First of all, I mean, I'll just say, I agree with you. That is a difficult thing to measure. I think that when we think about things we can try to control when it comes to security, it is things like exposure. We try to reduce the impact. I think we think about impact reduction versus being able to measure what the impact was on an individual, because I think that's hard to do as well. I think that understanding whether or not, for example, whether or not a narrative had actual impact on people in a community would take time to understand, because I think it would have to stem from a real study of what type of engagement happens following. Was there more agreement or uploading or was there more of the sentiment present?

Maybe those are things that could be looked at. But I think right now, in the space that we're in, the goal is really to reduce the impact of such things by reducing the presence of those things, by reducing exposure of those things. If we're able to see it in as real time as possible, and then take action to reduce exposure to content we know is put there with malicious intent, then I think we're doing the job of not needing to necessarily get to the point of assessing impact because we've already taken away the greatest piece of that.

Luke McNamara:

Right. I would imagine when you think about one way to approach the prioritization of those threats, particularly in the information operations space, looking at things like where you have a user base, or thinking about different regional threats or different communities that you've seen targeted in the past, those would all play into how you might focus resources on addressing sort of emerging threats that are coming out on the platform. Because as you noted before that it's a very dynamic space. And so you're not necessarily dealing with the same known threats or entities when you're initially looking at those reports of that activity emerging.

Aylea Baldwin:

Yeah. I think that for us it is very important to focus on where we exist as a platform. Because those are obviously the populations that are susceptible to these operations, to these malicious activities. So in terms of how we think about prioritizing, for sure one of the things that is top of mind is do we have communities that are relevant to that activity? Is it because of geography? Is it because of a certain concept or ideology or all of these things that can be represented as communities on the platform? Do we have a presence there and even geographically in terms of like literally geographically, is it related to a place where Reddit has a large enough user base and presence for it to have significant impact? So I think for sure those are things that we evaluate in terms of looking at which things need to be moved up on the list or not just based on the potential for impact to the communities.

Luke McNamara:

One thing I want to go back to you mentioned that I realized I should have followed up on, but you were talking about how awareness of the sort of drivers of information operations or other malicious state behavior are often shaped around geopolitical events, things happening in the news. How do you stay on top of what are those things going on? Just from an information standpoint, I imagine you're pretty active, at least in following some of the current memes and culture on Reddit to understand that component. What sources of information do you go to stay ahead or imagine where these threats might lead?

Aylea Baldwin:

I definitely consume a lot of news, just general world news. Also I participate in groups that are industry level having discussions about things that are going on, things that other organizations are seeing and trying to kind of see if that's something that we also see happening on Reddit or see being talked about on Reddit. Also just being on Reddit, honestly, understanding from being a participant in the community of communities, knowing what is the issue of the day.

And it's funny that you mentioned memes because that has also been a fun learning experience for me. Because I think I had a surface level understanding of memes before and once you really dive into Reddit, you really understand how much of a cultural thing they are and how much you really can learn about what's going on in the world just by understanding what's going on in meme culture. Yeah. So I think those are some of the things that I try to keep on top of to just have a general awareness what is going on and what could be bubbling up as a potential hot topic or something that could be viable, let's say, as a vector for someone to kind of take hold of it and use in malicious operations.

Luke McNamara:

So I have one final question for you, but before that, because you talked some more about means. GIF, how do you pronounce that? It's the eternal debate, right?

Aylea Baldwin:

No, and you know what, I don't know if I should say this wit shame, but I like to say GIF. I know it's GIF, I know it is.

Luke McNamara:

That was the wrong answer. But some people would disagree with me on that. I guess, and this picks up off of what you were talking about in terms of the different places you go to get news, and then pairing that with what you're seeing that's taking place on platforms. But where do you see, and this is a very, very broad question so feel free to answer this however you want. But where do you see disinformation going? As you noted, it's dynamic and the drivers are going to change. Some of those will be constant. And I think some of those things we can imagine the types of threat activity we can see emerge around particular things. But either from a TTP standpoint or things that will shape and drive disinformation, where do you see that going in the future?

Aylea Baldwin:

I hope will happen is that, like I was saying earlier, we would as a security community and as cyber threat intel in particular inside that community really put a lot of emphasis in the idea of getting more foundational understanding of information operations across the board as a standard piece of cyber threat intelligence. Understanding the psychology of these things, understanding how actors have leveraged this as a tool and having case studies and scenarios that are apparent to bring forward a new kind of thought process around this as a main staple in cyber threat intelligence analysis in particular. So, yeah, that's kind of my idea on where I hope it will go. I don't know if it's a prediction, but definitely a thought about future direction in this area.

Luke McNamara:

And do you think that we will see, I mean, one of the pieces that there's always been discussion around has been the usage of deep fake technology, artificial intelligence manipulated imagery. And I think that's for the most part what we've seen when we had Lee on last year, he was talking about that, a lot of it's been manipulated still images, but I think the concern is always that we'll see videos being utilized and targeted increasingly more. Do you think that we'll see more of those technologies as we go into the future being used for disinformation?

Aylea Baldwin:

Yeah. I definitely think there's room for that to become more of an issue as actors seek to understand more about how to leverage it and how to effectively leverage it, I think is key. So it's something that we should most certainly have on our radar and try to get ahead of, but it's going to be hard, because that's one of those things that is, from what I can understand today, is not something that's easily detected.

So I think that, yes, we will likely have to pay more attention to this as a real threat if the threat landscape is evolving as always and new technologies always present new dangers, new threats. As actors think about how to leverage this effectively, we may see more campaigns that involve these things. So I think it's not necessarily that I'm saying I'm predicting more of it, but I am certainly saying that there's a strong possibility that as this becomes a tool that is more attractive and there is more discovery for effectiveness of its use, that it's something that we will most certainly have to get ahead of and have on our radar and figure out ways as a community to try to solve for.

Luke McNamara:

And I guess in some respects that's no different than the always evolution between the defender and the attacker where innovations on one side for detection and response and then the attacker, the adversary, comes up with a new way of exploiting communities, discussion, networks. So yeah, I think it'll be interesting to see where it develops as well. It'd be fascinating to follow. Aylea, as always, great to talk to you and thanks for coming on and sharing your insights into what you're doing over at Reddit.

Aylea Baldwin:

Thanks for having me. It was really great.

Luke McNamara:

Take care.