The Safety Tech Podcast [Episode 2] - “I went down the rabbit hole”: exploring the real-life effects of false narratives

Tue, 4/19 11:24AM • 37:00

SUMMARY KEYWORDS

disinformation, disinfo, misinformation, content, people, narrative, sites, platforms, vaccine, organisations, mis, anti vax, models, advertisers, thinks, world, information, safety, network, breitbart

SPEAKERS

Heather Simpson, Ben Whitelaw, Clare Melford, Lyric Jain

Ben Whitelaw 00:06
Welcome to the safety tech podcast brought to you by the safety tech innovation network. My name is Ben Whitelaw, and I'm the founder and editor of everything in moderation, a weekly newsletter dedicated to online safety. Today we'll be exploring a topic that almost everyone has experienced themselves in some form, mis and disinformation. Although fabricated information is by no means new, modern technology has caused false narratives to proliferate at extraordinary speed and scale in recent years. That's scary enough as it is. But we know that unchecked claims and conspiracy theories can also fundamentally alter the way that people see the world and act within it. But while technology has played a role in amplifying misinformation, and manipulating public opinion, it could also be key to finding a solution to this growing problem. Thanks for joining us.

Heather Simpson 01:12
I was obsessed with Facebook. I remember in college, I was known as the girl that just constantly posted. So I used it, you know, for community, to be social to make friends, all that stuff.

Ben Whitelaw 01:28
This is Heather Simpson from Dallas, Texas.

Heather Simpson 01:31
When I was thinking about having a child with my then husband, we started looking into all the things that a parent does, including vaccines. And that's when I stumbled upon a nine hour Docu series, just detailing how scary vaccines are. I mean, it blamed cancer, autoimmune diseases, autism, everything was blamed on vaccines. And it was doctor after doctor after doctor. The Docu series was presented as an ad on Facebook. You know, the internet basically read your mind when you're thinking something

and things have trade time. And after that, I just went down the rabbit hole, I found an entire community on Facebook of anti vaxxers. That just reinforced what I believed and it was just downhill from there.

02:20
Simpson received 1000s of Facebook friend requests from other anti Vax mums and her posts began reaching 1000s of people on the platform.

02:28
It became kind of like my entire world. And it was I mean, I loved it. And just having that validation made me feel like I was doing something important and right. And then I actually started getting a lot of hate. And there were all these pages formed against me. I did not handle the hate. Well, I would wake up at 3am and check my phone and then I would just like start shaking, I was so freaked out because I would wake up to 5, 10 thousand hate comments.

Ben Whitelaw 02:56
But some of the things that people were saying started making her question her beliefs.

03:00
I had to get endometriosis surgery a month before the pandemic and my friends were saying, my community was saying you need to eat healthier. You know, surgery is the easy way out. Like you have endometriosis. Because of your diet, you just need to eat better, and you wouldn't have that. And so I was upset. I just didn't like the shame culture I was experiencing. COVID hit around March / February of 2020. So that next December, I was talking to a pro vaccine friend, I had spent the last few months reading more pro science pro vaccine doctors and material and I just talked to her and I was like, you know, I'm not for every single vaccine yet, but I'm for some, and she was like, You are pro vaccine, if you believe that they are safe and effective. And I was like, wow, at that point, I just went for it. You know, I had so much anxiety, vaccinating myself and my daughter, but with every shot I got that fear just went away and I came more and more confident each time. I was just so relieved that I got vaccinated. I mean, my daughter's too young to get the vaccine right now she's only four but just to help protect her and to help protect her from losing her mommy. I was just so relieved. Seeing my friend's parents die was just devastating. And it just reinforced to me that I definitely made the right decision.

04:27
Since getting vaccinated, Simpson has co founded back to the vax to convince members of the community that she used to be part of the vaccines are safe. Just last week, she was contacted by an old friend of hers.

Heather Simpson 04:40
My friend that used to be on my friends list, she can testify to watching the mass exodus from my friends list and watching you know, all of it hit the fan. I fell out of touch with her when I was excommunicated from the anti Vax world, and she remained an anti Vaxxer up until Two weeks ago, she is now an influencer on Tik Tok. And she's been seeing more and more pro science videos and she called me just completely panicked, saying I think I'm a pro Vaxxer now, and she thought her COVID shot she scheduled her kids first vaccines as well. And the worldwide group effort of people that have

changed, and people that are pro vaccine, making a graceful effort to persuade others it is working, these little seeds are being planted. And it's kind of this group effort. And we're seeing the fruit of that.

Lyric Jain 05:35
In general, we view individuals who believe in misinformation campaigns as individuals who have been targeted.

Ben Whitelaw 05:41
Lyric Jain believes it's important to note that people like Heather who have spread misinformation are themselves often victims of false narratives.

Lyric Jain 05:49
There's a world where people are being engineered to have certain opinions and certain viewpoints, where these are being engineered either by known domestic operatives that have a specific agenda, or operatives of foreign government, and I think one can view them through a lens of them being a victim of how this disinformation and misinformation campaigns work.

06:10
Jain is the founder and CEO of safety tech firm Logically, a company working with platforms and governments around the world to tackle mis and disinformation.

06:19
We bring together expert intelligence and artificial intelligence to tackle mis info and disinfo. So everything from misinfo that might be present in content to how disinformation campaigns are engineered to influence populations around the world, we try and identify those with speed and scale, and help our partners across the public sector, particularly in areas where missing this info might be impacting public health, public safety, election integrity and national security. And outside of that, we also work with platforms to help them implement kind of scalable content moderation systems that identify misinfo and disinfo.

Ben Whitelaw 07:01
How do you define misinformation

07:05
Into a thorny one right away. I think this is something that there's been a lot of consensus around finally across across the industry. So Mis info is really one that the area that's focused on kind of false information and falsehoods that are contained within content, but there's a larger bucket that perhaps is more problematic of disinformation where, regardless of what the content saying, maybe it could even be true. So literally, it's a campaign that's been engineered with elements of inauthenticity, say bots, or a degree of deception to try and influence populations in some way

Ben Whitelaw 07:40
in your work, or logically, what kinds of impacts are you seeing from mis and disinformation

07:45
it's unfortunate that we see a lot of problematic events happening every day around the world because of misinfo and disinfo. But there's also kind of a chronic kind of breakdown, that's happening as well as the kinds of events we see happening are kind of the big, almost security threats that organisations and countries around the world are faced with. We are seeing that with previous incidents in the UK where kind of 5G towers have been burned down vaccination centres have also been attacked. Threats to life have been made to kind of people that are trying to be impactful work in vaccine rollouts equally, people are trying to help tackle climate change just because there's these conspiracy driven communities that are threatening law and order, but also the lives safety and the operations of these critical public sector functions.

Ben Whitelaw 08:32
And tell us a bit more about logically and how it came into being in the first place.

Lyric Jain 08:38
It almost feels like ancient history now but unfortunately, a series of strange events - had a bit of family tragedy in 2015 / 14. My grandma, she was 85 at the time, but she still used WhatsApp. She got a tonne of these messages saying, hey, drink this special green juice, give up your cancer meds and you'll live longer. And unfortunately, we lost her a lot earlier than we ought to have. At that time, I didn't really put it together as misinfo disinfo just thought it was lone wolf fraudulent activity. Started going down a rabbit hole of what echo chambers were online, pre the academic interest in the space, particularly around the European referendum, I think my experience was quite novel and my hometown Stone happens to be the highest Brexit voting town in all the UK and where I was, at the time Cambridge was the highest remain voting town of the UK. So I kind of vividly recall this memory to my friends from both those places who were in town the same day and their social feeds were completely different by talking about the referendum, completely different information, both in fairness with degrees of misinformation. And that felt like there was something there. And it was that kind of observation, missing kind of my academic interests at the time working at CSAIL in the Media Lab at MIT that led to the initial genesis of the idea of hey, maybe that's something that automation alone could do here and quickly hack together something that went out there and identifies misinformation on one social media platform in particular, and we were like, Hey, we're doing that we're doing a better job than they were at that time in 2017. And that proof point was kind of enough for me to kind of build a company around, we're now 150 ish people split across the UK, US and India. Proud to be supporting kind of the UK gave US gov India gov with the challenges around misinfo and disinfo. But also, Facebook Insta and Tik Tok. So, lots of traction that we're proud of already. There's still an awful lot left for us to do in helping those stakeholders, but also, so many other organisations that face similar challenges with mis and disinfo.

Ben Whitelaw 10:45
Data ethics is a very hot topic at the moment. And I'm curious to know how logically builds data models that are, you know, ethical and inclusive. And, you know, we've had a number of controversies recently where models are built by humans who live in San Francisco and aren't very representative of the users of the platform. How do you address some of those challenges?

Lyric Jain 11:14
I think it's super important for us to be explainable. In particular, when we're calling out something as being potentially misinformation or disinformation. What we mean by that is, we can't just say, hey, we believe this is misinfo, because the model says so. There's a little anecdote, I believe this is still true. It was certainly true a year ago. And it's if you went to the valley, you saw a self driving car driving about, and it suddenly crosses a red light, and a policeman pulls it over and asked the driver, Hey, why do you cross red light? Driver says I dunno the car did it. The policeman asks the car the question, the car would respond, I don't know. And the aim with explainability both to that domain as well as to the misinfo disinfo domain would be to have the model answer, why it's making assessments. And the reason it can't, again, in self driving an example, for instance, is particularly because they rely mainly on really large blackbox models that historically haven't been compatible with explainability. The kinds of models that have been compatible have been traditional statistical learning, statistical machine learning models, those recent innovations have been made explainable, that through kind of more recent innovations, even these big blackbox models are becoming more interpretable. And there's a lot of research that the AI community in general is doing on this, that we're specifically focusing on, how do we make our inferences on misinfo also explainable? So it would be things like, hey, is this our credibility system that said it or our veracity system and our veracity system would be taking a claim like Ben says, Lyric doesn't like orange juice and breaks that down into Hey, do we fact check the fact that Ben said that or the fact that Lyric doesn't like orange juice and kind of it would specifically call out the appropriate bits of evidence that we used, how much different pieces of evidence from different sources were wasted? All of that would be would be visible. So those are the kinds of investments we're making in terms of making all of that possible. But generally, given the nature of our work, the organisations that we work with wouldn't trust us if we didn't have that level of rigour anyway. So it's almost a qualifier for anyone working in the safety tech space to to have a degree of explainability baked in. The only caveat to that might apply to kind of super sensitive cases where people might not care how you got the answer, but you you just need a super accurate answer all the time. But for misinfo and disinfo just because of the broader public ramifications of it, everything has to be explainable. And that's kind of the logic that we always follow.

Ben Whitelaw 13:39
That's really interesting. I guess one area that we're seeing a great deal of misinformation is closed networks, Telegram being a great example. And I wanted to get a sense from you what the specific challenges are when it comes to those networks.

Lyric Jain 13:54
Yeah this is an interesting one for us. Because we've been on this one for quite some time. Because some of the work that we do is with elections in India, we noticed this trend back in 2019. That increasingly a lot of misinformation already, just because of closed network adoption over there, particularly WhatsApp was just dominant. So the challenge there is kind of multifaceted because a if it's a fully private setting, there's no way in which the platform can moderate that, no way in which anyone can do a lot about it unless we use quite intrusive methods, which would be legal in most countries, or which would have huge concerns around user privacy and platforms making choices around how much privacy to enable for free for users. But that doesn't mean that there's nothing we can do. Because one of the phenomena we've seen around closed networks is that people who tend to get banned on other

platforms either because they're known extremists - those individuals and groups tend to coordinate on these platforms, and increasingly now that even these conspiracy actors have been running from various platforms like followers of QAnon, there's also moving into those groups and going through this rabbit hole of radicalization where they're getting further radicalised, the more they engage in those groups and the kind of more extreme groups, if you will, kind of the full on far right groups, those are using this as effectively the ground for recruitment to make sure that their own movements and interests get fed permitted. So there's something quite interesting about the way in which the Titan v net community can organise a pretty large campaign that doesn't just have an impact on telegram or WhatsApp, or this closed network that can plan and coordinate activity and viral events that they want to happen on other platforms via Twitter, or Facebook or an Insta, but also plan out actual real world events like, hey, I want to go and set fire to that vaccine centre, for instance. So this is getting quite hard to track them are also quite hard to do something about apart from just reporting them to the relevant law enforcement authorities being what we need is kind of greater participation from platforms that do have a closed network to invest in other methods to tackle closed network misinfo and disinfo. And we actually have a case study on that, on our website on some of the tactics that were used, particularly during the Indian election on WhatsApp. What we found was if we were able to fact check something and respond to a user within 30 minutes, I think it was maybe an hour, it would then be shared by that user and seen at least like 10 to 20 other users within the same network. The way in which you can get a fact track to the exact population that has seen the original piece of info so there's ways in which we can use the same mechanics that make a closed network so powerful for a bad actor, actually powerful for initiatives like fact checking. So again, still very early stage, it's not kind of peer reviewed research at this stage. So we need to continue across organisations such as ours, civil society, academia and platforms. Yeah, invest in those initiatives just to find finance because there's no silver bullet today apart from investing the types of monitoring operations that we have.

Ben Whitelaw 17:09
While Jain and Logically worked to identify false information. There are others trying to disrupt the business model that allows people to profit from online mis and disinformation. One of those people is Clare Melford, the co founder and CEO of the global disinformation index.

Clare Melford 17:27
So GDI is a nonprofit and not for profit, which we started in 2018, to disrupt the funding to disinformation websites. So we do it by assessing news websites, or on their risk of carrying disinformation. And we use a combination of human assessors and also artificial intelligence to assess the risk of a news site carrying this information. We then provide those risk ratings of news sites to the advertising technology sector and to advertisers themselves. So they can choose whether or not their ads end up being shown on websites that have a high risk of disinformation. So they can simultaneously defund the high risk the disinforming sites, and redirect their funding towards lower risk higher quality news sites.

Ben Whitelaw 18:24
And can you give a little bit of background as to how the global disinformation index came about?

Clare Melford 18:33

So GDI was conceived in the wake of both the Brexit vote in the UK and the 2016 election of Donald Trump in the US, my co founder and I, He's based in the US, I'm based in the UK. Although we didn't know each other had similar responses to those events in our countries that a large percentage of the population had been exposed to a very degraded information environment, and that may have led to a less well informed populace making decisions that ultimately are not in the economic interest of the country in which they live. So we both have some similar ideas on different sides of the Atlantic, we both wanted to see how we could address the business model that we saw, enabling and accelerating disinformation around the world. We both alighted on the fact that what was required in order to break the automatic link between engaging content clicks and advertising dollars was a way of assessing that content, assessing those news sites in particular, on their risk of carrying disinformation. And if you can neutrally and independently assess the risk of disinformation. You can give advertisers a choice about whether or not they want their ads on that site. But prior to GDI existing, there has been no way for advertisers to avoid their brands ending up on high risk sites, which creates a great brand safety challenge for them. Nobody wants their brand of toothpaste or shampoo to end up next to a piece of anti semitic content or misogynistic content or COVID disinformation, for example.

Ben Whitelaw 20:30
As far as that kind of rating goes, that you provide advertisers, can you talk about some of the signals that go into understanding whether a new site is is likely to carry misinformation.

Clare Melford 20:45
So the way GDI thinks about disinformation is through the lens of what we call adversarial narratives. So disinformation is not an event. It's a process that happens over time. And it happens through repeated exposure to content that is setting you the reader against the subject of that content. Its content that is creating an adversarial narrative about a particular subject. And those subjects broadly fall into three groups. The first is groups of people. So it could be content that is creating an adversarial relationship between you as the reader and people of a certain gender or colour or ethnicity, or religion, or sexuality. So the sort of the hate speech groups you might call them. The second category of adversarial narrative tends to be around what we loosely call institutions of society. So content that is creating an adversarial narrative against the media, the judiciary, the police, a democratically elected government, for example. And then the third category of adversarial narrative is against science itself. And we see this a lot with COVID disinformation with anti Vax disinformation with climate change denial with 5g conspiracies. And crucially, all three of these genres of adversarial narrative carry a very clear risk of harm, they all undermine trust in democratic societies, and they all carry a risk of real world harm. And in our view, that adversarial narrative lens actually gives you a much more nuanced and useful way of defining which content is actually harmful, particularly to a brand's image than simply a true false binary. In you know, a few years ago, people used to talk about fake news. But disinformation isn't about fake news, the harm isn't about whether something is true or false. If it was, we'd be trying to remove all mentions of Santa Claus from the internet. And we're not. Similarly, there are news sites, such as Breitbart, which carries a section called immigrant crime, which is a curated list of stories, all of which are probably perfectly true about crimes committed by immigrants to the United States. So the problem with that section of Breitbart website is not the truth or the falsehood. It's the narrative that's created. It's creating an adversarial narrative against people who are not born in the US with a false narrative that those people are more likely to commit crime than native born Americans, which isn't

true. So the true false dichotomy is not a helpful way to think about disinformation. The adversarial narrative lens is a more useful and nuanced way. And it allows advertisers to have much more control over the sort of content that they want their ads to end up supporting.

Ben Whitelaw 24:09
Is there any kind of specific examples of new sites that you know have been kind of defunded in the way that you talked about?

Clare Melford 24:18
I can only talk about the ones that are already in the public domain, which are not necessarily the result of our work, but of the whole ecosystem of organisations that are trying to do the same. Both of which are there have been many news stories about so the first is Breitbart. There's even video footage of Steve Bannon talking about how the advertiser - a reaction against the content on Breitbart cost them 90% of their ad revenue. And the second is the Gateway Pundit. There was a documentary on French TV two, I think in France last year, which showed a lot of the content on Gateway Pundit and talked to the advertising technology companies that monetize it, notably Google, and led to those companies deciding not to provide services to the Gateway pundit in future.

Ben Whitelaw 25:10
What type of misinformation? are you most concerned about in the work that you and GDI do?

Clare Melford 25:18
I think I am increasingly concerned about the not an individual narrative, but the meta narrative that democracy isn't working. Democracy is the best form of government governance human beings have yet invented as a way of enriching and developing stable, prosperous, peaceful societies. That is under attack. And it's been under attack for several years. But the year 2021 started out with a great democracy attacking itself with the capital riot on January the sixth. And that has continued throughout 2021. And if you look at a whole range of disinformation narratives, they are becoming, they are sort of blending together in support of this uber narrative that democracy isn't working, that there's a corrupt elite working only for themselves, not serving the will of the common man, and that only a strong man can get rid of the corrupt elite. So I'll give you the example of Q anon which started many years ago as a narrative around Deep State elites with some crazy stuff around these elites being paedophiles and cannibals and had a very strong anti semitic element to it. But it was niche. It then over the course of 2019 2020 2021 morphed into a narrative around COVID and anti Vax content, mask protests locked down protests, voter fraud around the 2020 election. And now it's even morphed into a narrative around abortion. So many of these disinformation, these adversarial narratives are blending and merging into the service of this uber narrative that democracy isn't working. And when that narrative takes hold, you get the events of this week - Putin has spent or the Russian State has spent years dripping into the Russian information ecosystem that Russian speakers in Ukraine are subject to genocide, that Ukrainian politics is based on a Nazi ideology. And that disinformation has softened up the Russian population to war, which is now what has happened. So there isn't a particular type of extremism or health misinformation or election fraud, disinformation that concerns me. What concerns me most is that many of those different narratives are being co opted into this meta narrative, which is successfully

undermining people's faith in democracy as the best form of governance humans have created. And if people lose faith in democracy, autocracy follows,

Ben Whitelaw 28:21
I want to kind of kind of ask a bit more about this idea that, you know, has courted GDI about the defunding of content creators. And and you've talked about this, this idea as a possible solution to some of the information ecosystem problems that we face. Can you explain this in a bit more detail for us, please?

Clare Melford 28:42
The first thing I'll start out by saying is that freedom of speech is a crucial pillar of any free society. But there have always been limits on free speech, even in countries like the US. When your free speech impinges on my human rights, that's not allowed anywhere. So hate speech, for example, in many countries, when it carries the real risk of inciting physical harm to the individual that is against the law in lots of places, yelling fire in a crowded theatre is not deemed to be free speech. It's deemed to be highly dangerous to the to the rights of the people in that theatre. So, yes, freedom of speech is crucial. But no, it is not always unfettered. And in no one's constitution, is there a right to profit from your speech - that is not enshrined in law anywhere. The current debate around disinformation, the solutions to disinformation has been far too focused on content removal or content moderation. And that is problematic that there is some content that shouldn't be there, which we can all agree. Child abuse videos, pirated videos, there is content that is illegal, should be illegal and should be removed. But it's using content removal to deal with disinformation can be problematic especially in repressive regimes which can veer towards censorship. However, reducing the financial incentive to create harmful content is a much more market friendly democracy protecting way of nudging the information ecosystem towards a much healthier place. And there is a lot of financial incentive right now to create disinformation. It's not the only reason disinformation adversarial narratives are created. Some of that is ideological. I don't think Putin is spouting his lies around Ukraine because he thinks it'll make him money, he knows it's going to lose them a huge amount of money. So a lot of the political propaganda is often ideological. But the financial motivation is very clear. And it's entrenched in the way we have decided to structure the internet today. Because the internet and the services on the internet are largely free. They're paid for by advertising. And when there are no barriers to create content, on the internet, the only way to make your content stand out enough to get advertising dollars is to make it really stand out versus all of the other content. And the best way to do that is to tap into our negative emotions of fear or hate or greed or disgust or anger, because we're much more likely to respond as human beings to like to share to comment on content that triggers our negative emotions than we are that trigger our positive emotions. So that has embedded into the structure of the internet today, an incentive to race to the bottom to make the nastiest content possible, because that's what drives engagement. And that's what drives the money. What GDI does is distil down a list called our dynamic exclusion list, which is about 2000 sites of really the most disinforming sites on the Internet in multiple languages that we track all the time through our machine learning classifiers. That list is an exclusion list that we can provide to advertisers to enable them to make sure their ads don't end up on those sites. And we know that that's had a significant impact on the advertising revenue that has gone to those sites over the last 15 months, 18 months. So we know that that does actually work. And we can assume that those advertising dollars that are no longer or are to a lesser extent, on those highly disinforming sites, they

didn't go back into the pocket of the advertiser, they will have gone towards other sites on the Internet, and hopefully they will have gone to higher quality news journalism, which is a desperately needed pillar of free democratic societies.

Ben Whitelaw 33:03
If you think disinformation is a challenge nowadays, it's only going to get worse. Experts believe that fringe views and false facts are going to become an even bigger problem. catalysed by huge jumps forward in artificial intelligence, and computing power. I asked Claire what she thinks the future holds for our online information ecosystem.

Clare Melford 33:24
I think we will look back on this era, the sort of late 20 teens early 2020s as an anomally in years to come much as we look back now on smoking in aeroplanes or smoking in the tube and the London Underground as something where we can't believe that was ever allowed, because it was so clearly harmful to health and a virus. I think we will look back on this period as very, very light regulation coupled with very misaligned dissent incentives to create harmful content as a bit of an anomaly. I think we will see regulation that promotes competition across multiple sectors. And we will see regulation that establishes a minimum floor on the types of content that can be profited from. And I think we will see more of a coordinated and systematic approach to assessing risk of content, rating content on that risk and sharing that risk information across platforms. But establishing that sort of global systematic, shared resource is a huge governance challenge, that we are only beginning, we're only at the foothills of figuring out how that can work in a democracy protecting free speech supporting way.

Ben Whitelaw 35:04
While safety tech will certainly provide part of the answer, Claire believes that media literacy is an underused weapon in the battle against misinformation.

Clare Melford 35:13
There are four things that we would say need to happen in order to significantly reduce the risk of disinformation and its harms. The first is that you can identify disinformation risk in real time. The second is that there are tools to curtail the spread of that disinformation to push it down people's newsfeeds. The third is that the incentive to create it is reduced so you can disrupt the funding, which is what we've been talking about today. And the fourth is that the susceptibility to that disinformation is reduced. And that is the media literacy piece. So it's a hugely important I would, I would say, it's the fourth leg of solving this challenge. But it's the work of education, you need to start in school. And you need to train kids to be critical thinkers and to ask basic journalistic questions about sources, thinking about incentives, and journalistic freedom if people are trained to think even in a small way as journalists think, than that significantly helps.

Ben Whitelaw 36:19
If you'd like to hear more about the ways safety tech is tackling disinformation, head to the safety tech innovation network, an international network dedicated to the promotion, collaboration, and industrial application of online safety technologies. Become a member to receive the latest information about

safety tech events, and reminders about future episodes of the safety tech podcast. Thanks for joining me, and I'll see you next time. This has been a 4kicks production.