Center for Humane Technology | *Your Undivided Attention* Podcast
How Will AI Affect the 2024 Elections? with Renee DiResta and Carl Miller

Tristan Harris: Hey everyone, welcome to *Your Undivided Attention*. This is Tristan.

Aza Raskin: And this is Aza.

Tristan Harris: Sometimes it's really difficult to get a grasp on exactly how AI is going to impact our lives and our democracies.

Aza Raskin: And one of the ways Tristan and I like to explain it is that social media was our first contact with AI. As a society, we're now very familiar with all of those downsides of unregulated tech, and what we're starting to see now is our second contact with AI.

Tristan Harris: 2024 will be a massive global experiment in the potential for how our second contact with AI, that is generative AI, creation AI, can supercharge the harms of social media. 2024 will be the biggest election year in world history. There's something like 2 billion people who are be undergoing democratic elections this year from 70 countries, including some of the world's largest democracies, the United States, the UK, Indonesia, India, as well as countries like Taiwan, Brazil, Venezuela, Russia, South Africa, and Mexico.

Aza Raskin: So today on the podcast, we're going to be talking to two experts on how the new wave of AI is going to crash over democracies. There are too many elections for us to cover in one episode, and the experts we've selected focus on the US and the UK. But many of the ideas and trends we cover here apply globally.

Tristan Harris: Our guests today are Carl Miller, who is research director for the Centre for the Analysis of Social Media at the UK political think tank, Demos, and Renee DiResta, who's an old friend and technical research manager of the Stanford Internet Observatory, where her investigation into Russia's Internet Research Agency was highly influential to the Senate Intelligence Committee's findings about what Russia did during the 2016 elections. And she's been a guest on this podcast before. Welcome Carl and Renee.

Carl Miller: Hi there.

Renee DiResta: Thank you for having me.

Tristan Harris: What is different today if we're doing a situation assessment about the threat model from generative AI going into the elections this year that was not true four or five years ago. Renee, do you want to kick us off?

Renee DiResta: Sure. So there's a couple of things that have changed. First in the realm of social media itself, there is a proliferation of new entrants over the last four years. I call it the great decentralization. There's people who are moving to federated social media platforms. There are entrants like Threads or Bluesky or Mastodon. Mastodon has been around for a while, but I think again, people are migrating to

it and migrating away from Twitter. And it's not homogenous across all communities. But certain communities much the same way there was a proliferation of the creation of alternative social media platforms that catered to the interests of right-leaning users, you now see the same thing happening on the left. So you have more people across more platforms.

And then there's also the thing that I think we're going to talk the most about today, which is the additional impact of a new technology that layers on top of all of that, and that is the generative AI dynamics. So generative AI was available, but in a very limited sense in 2020. It was not as sophisticated as it is now, and I think far fewer people were aware of its potential in 2020. So you have the same way that social media took the cost of dissemination effectively to zero, generative AI has taken some very sophisticated content creation costs down to virtually zero. So you have the transformation in the social media ecosystem, that's shift number one.

Shift number two, you have increased polarization, increased tension, distrust within society. So that's a social problem, not a technical problem, but these two things intersect. And then the final piece is the layering in of generative AI. So a new technology that enables people to create unreality. So I think we have the intersection of these three major dynamics all coming together in one of the biggest election years that we've had in quite some time.

Aza Raskin: I think when people hear that, I wonder if they think, all right, so the problem is going to be more and wider distributed, mis- and disinformation. But Carl, I want to turn to you because I know that you've been thinking about beyond just more false information on reality, there are deeper risks that emerge with generative AI. So I'd love for you to talk about that.

Carl Miller: Yeah. Deeper risks indeed. But actually before I dive into the balmy waters of AI or really any technologically driven change at all, I do actually want to dwell for a moment on the actual conceptual development of information warfare and influence operations, because I think that's as important. And I think it's important to note at the beginning that what we're dealing with here, and at least the kinds of online maneuver that are going to be most injurious and damaging to elections specifically won't just be disinformation being spammed around the internet. These are going to be coordinated, concerted, evaluated, measured, and funded campaigns of one kind or another. And what's underlying that is both a trade craft and a mindset. So it's a mindset that sees information as a theater of war. And I think that is a fairly novel conceptual pivot actually.

I think if you go back to the '80s or the '90s, you don't actually hear about militaries so much talking about information as a space that they need to dominate. It's much more considered a tool or even a weapon, but not up there with air, sea, land and space itself as a theater of war. I think that's really important. But then the trade craft, like how should information be competed

within? What does the struggle look like? How should our strategy be composed? And I think that actually from 2020 through to now has changed quite a lot as well. I think we are increasingly seeing the deployment of cognitive psychology and behavioral science. So more sophisticated understandings about how influence works and what happens when you surround people by different kinds of information. And also I think increasingly campaigns which join up lots of different kinds of influence together. And I think we really do run the risk of thinking this is a social media centric or social media only phenomenon. It absolutely isn't.

If we're talking about state bureaucracies, states at all, or even sophisticated private sector actors here, they're using economic inducements. They might be using coercive means, they might well be putting assets down on the ground, they'll be using people, they'll be bribing people. They'll be using all kinds of ways in order to achieve geopolitical advantage through an election or reap the influence that they want. So I actually think just before we start talking about artificial intelligence, it's important to note that it's quite likely, at least what I see, that the actual ideas that are informing the kinds of exploitation of tech and information maneuver, I think are getting more subtle, more rarefied, and really better informed by this weird grab bag of different applied academic disciplines that they're looking at.

Tristan Harris: What I didn't understand in that Carl was the who behind some of those examples. So who has this increased cognitive knowledge and doing the economic inducements and the other things that you mentioned? Just to be clear.

Carl Miller: Yeah. Thanks. And I think this touches on Renee's point around that the threat actors actually who they are have become more diverse probably. I mean, one of the big trends we've seen since 2020 is the emergence of a for-profit series of offerings. And they're spread across the light net and the dark net. We see the shopfronts. We know some of the companies. Some of them openly operate in Europe, some of them not. But it's likely now that we are dealing with state bureaucracies, be they military or otherwise. We're dealing with for-profit actors, we're dealing with political campaigns and we're dealing with consultants and smaller actors as well.

Renee DiResta: So I think maybe I can talk a little bit about that. So at SIO, we have assessed influence operations internationally since 2019. So a very broad swath of actors. And while a lot of the focus is really zeroes in on the American culture war, the American political polarization, that notion of actors expanding into for-profit enterprises. A lot of what we've seen, for example in the Middle East operations being run out of Egypt are run by what we call digital mercenaries. And the mercenaries are entities that are for hire. Oftentimes they're social media managers. Oftentimes they actually manage the accounts of very legitimate people.

So sometimes you'll see when Twitter or Facebook, Twitter in particular would take down a network, oftentimes other real clients of the company that was doing malicious things would also temporarily lose their accounts as well and then have to file to get them back. And that's because Twitter just took down anything associated with the network that it was disrupting.

Tristan Harris: Let's quickly talk about deep fakes, because this is what's getting the most attention in the press and from politicians. And in September right before the Slovakian election, there was a deep fake audio recording of a political leader seemingly plotting corruption two days before a very tight election. And his opponent went on to win. A week later there was a viral deepfake of UK labor leader Keir Starmer on the first day of the Labor Party conference. There's no question that this is happening more, but do we have any way of accounting for how much influence these videos and recordings are actually having?

Carl Miller: Yeah. Deepfakes, generative images and video, they seem to be the most straightforward way that AI might change illicit influence operations. But actually I think it's pretty incremental. On the one hand, we've been able to fake videos and images for a very long time. I mean, you go to any proper production house anywhere in the world, and you'll see lots of examples of this happening way before AI came along. And it probably isn't really how influence often works. I mean, we know that influence often flows through the social connections which join us. It has to do with meaning and identity, how people feel as much as how they think. So AI might change incrementally the use of fake images. It might make it cheaper, it might make them slightly more convincing. But I don't think that's really what's going to change the game.

Aza Raskin: Renee, President Biden wants US agencies to start watermarking content that AI has generated. So that we're clear for listeners what watermarking means is that it's essentially that digital content would have some mark in them that would let your computer know that this was AI generated. And I want to ask you, I'm pretty skeptical of this, but I want to ask you, what is your faith in this solution?

Renee DiResta: Well, I think it's an important thing to do. So Dave Wilner, formerly of OpenAI trust and safety is now a fellow at Stanford Cyber Policy Center. And we wrote a thing about this in Tech Policy Press for anyone who wants to read the long details. But suffice it to say, I think that the executive order adds a government imprimatur to what has been actually an ongoing industry effort. So watermarking and provenance has been something that a lot of companies have been talking about over time as they've tried to figure out the question of how do you revise your synthetic media policies? Social media companies, for example, know that even though the creation is not necessarily happening on their platform, they're going to be the distribution vector. So they're very interested in this question, they're participants in the conversation. And if you have watermarked content like a machine-readable watermark, that's the thing where the platform might decide to signal that the content is AI generated.

But this does assume in adversarial models and in adversarial spaces, you're not going to have good guys using watermarked content. And that's because even if the majority of the large public providers where the average person can go and use an interface provided by OpenAI or something, that piece of content might come out watermarked. But if you use open source models, it will not. And one of the areas that SIO has spent a lot of time working in this year has actually been the rise of AI generated non-consensual intimate imagery and AI generated child exploitation content. So that's been where a lot of our team focus has been this year. And what we see and what we've written about and what we talk about actually is even as people focus on watermarking and election integrity, the egregious things that are happening with some of these models and other spaces are really extraordinary.

So I think the challenge with watermarking is you're going to have an intermediate period where in addition to the dynamics of good guys using them and bad guys using open source tools to oversimplify the statement. What you're also going to have is this question of what happens when you have content that is taken on a phone and edited slightly on a phone? Where does that fall? So there's just the notion of content provenance I think is very interesting space, very evolving space, something I think that is very important. But it is not a panacea for addressing something like malicious deployment of AI generated content into either something like election narratives or market manipulation tactics or non-consensual intimate imagery. So it's useful but not the solution to all the problems people are concerned about.

Aza Raskin: Carl, can you talk about how you see the weaponization of friendship playing out as we're heading into the 2024 elections?

Carl Miller: Yes. So I think a lot of the applications of artificial intelligence feel fairly incremental. We've already been able to manipulate videos for quite some time. I mean, anyone just has to go to Hollywood to know how effective that can be without touching artificial intelligence. Likewise, the creation of backstop identities online, it's been possible for a long time. There are annoying bits. There are probably some things that be made easier using AI there too. But the one that really keeps me up at night, and I don't have any evidence this is being used, but I would be astonished if this isn't being explored is not trying to send messages to a very large group, but instead trying to influence a target audience by establishing a great many direct one-to-one relationships with that audience.

If we know anything about how influence works, we know that it spreads down social ties. We know that enduring senses of kinship and belonging and meaning and friendship, these are the things that really change people. Not being spammed out by some anonymous account online. And I think to me, the game-changing application for AI and illicit influence would be to actually now power a whole series of either automated or semi-automated friendships between you as the influence agent and that target audience. And you can just

imagine they'd always be there ready to lend an ear. Always there ready to ask you how your day was. Ready to sympathize with the things that went wrong in your day. Ready to celebrate your successes. They could be like the perfect friend.

And over time in a way, which I think would be extremely difficult for people to detect, I think almost impossible for researchers like me, you could just begin to use those relationships to suggest ideas, issue salience, making sure they've seen certain stories coming up, certain controversies. It could be extremely subtle and long-term, and swimming with people's cognitive biases and swimming with all the ways in which we know human beings work and the heuristics that they have. So I haven't seen this, and I don't know if I ever could. I don't know if researchers ever could see this. But that to me is how AI might completely change the way that influence works.

Tristan Harris: Yeah. What this makes me think about is there's the cost of distribution, which has gone down because of social media. Then there's the cost of content generation, which has gone down because of AI. But the other thing I hear you saying is that the cost of one-on-one fake friendships, which can use as a vector for fake influence, that that is going down to zero. And I want to mention, I actually know researchers who have a bunch of fellows who are experimenting with what are the worst stuff you can do with generative AI, and a friend has a 16-year-old intern who used GPT-4 API to create a discord bot that starts striking up relationships with people in discord. Basically taking keywords of things that they're interested in, like astrophysics or whatever, building a little friendship relationship with them. And then you can start sending them other news and articles to say, "Hey, check out this." This is a 16-year-old programming this little bot. If a 16-year-old can do that, imagine the kinds of things that we're really stepping into.

And to your point earlier, Renee, instead of just social media platforms, we also have many smaller group platforms, Discord, Twitch, I'm sure Telegram. You could list many others. And that's different from 2020 where there was slightly more concentration among a handful of platforms. And Aza, this reminds me of something that you said in our AI Dilemma presentation, which is that loneliness might be our biggest national security threat.

Aza Raskin: Yeah. That's exactly right. I want everyone in the audience to scan their mind for the times in their life that they've most changed. How did that change come about? And I'd argue as you scan your mind, most of them have come through a relationship, maybe through a parent or a best friend or a girlfriend or a boyfriend. It is those people that we encounter upon our life paths that change us the most and irrevocably. So we are in a sense outsourcing what I'm hearing Carl say, humanity's most powerful and most influential technology, which are relationships. And all of a sudden, as you were saying, just on the cost of generating relationships, intimate relationships drops to zero. And what Carl is

saying, which I think is fascinating, is that it could be happening now and how would we tell?

Tristan Harris: Yeah. That's such a fascinating part. Renee, I wanted to quickly let you speak to this because I know that in 2020, 2016, there was some things that we saw certain actors do with building one-on-one relationships with users as well. So I want to quickly give the audience that evidence point and then talk about how that threat's going to keep evolving.

Renee DiResta: So the research that I did for the Senate Intelligence Committee, the data sets were provided by Twitter and Facebook and Alphabet. And one of the interesting things that we did not get to see was the engagements that were done over Messenger. So we knew that they were doing them because one of the things that the Internet Research Agency constantly did in this Facebook pages, not so much on Twitter, but constantly on Facebook, was put out calls to connect. So they were constantly saying, "Hey, are you a designer? Are you a photographer? Are you a this? Are you a that? Slide into our DMs? We want to talk to you about a project we want to hire you for." And sometimes, oftentimes that was like photographing a protest or helping to support a protest.

Tristan Harris: I just want to jump in here and remind people that the IRA, which is the Internet Research Agency, was the troll farm in Russia that interfered in the 2016 presidential elections. It was run by Yevgeny Prigozhin of the Wagner Group, which had very deep ties to Russia's intelligence services.

Renee DiResta: So one of the things that they would do is they would engage with activists who of course at the time were not thinking, oh, this is a Russian. Some of them were suspicious. You actually do hear some of the Black Lives Matter activists who did in fact engage in direct messaging with these folks noting that some of them did say that they felt that there were red flags. Like something about the communication was off in some way. But they did again, even as far back as the 2015 to 2018 timeframe with these datasets, we couldn't see for privacy reasons the specific types of relationships that they formed. But what we could see was these constant exhortations to engage in those relationships. The offering of let us help support you, let us provide you with posters for your protest. Let us provide you with connections with funding and what is it you need. And they position themselves as very helpful.

So, again, I think as Carl notes, this is not unique to the age of the internet. This is how influence operations and agents of influence have conducted themselves long before social media was a thing. But what it does is it makes it easier, because you don't have to see the person face to face to have that interaction. You don't have to talk to them on the phone. Certain other tells that might be visible are not quite so visible when you're engaging in a chat relationship. And also as Carl notes, that peer-to-peer friendship is again the thing that is consistently shown to be the most influential. And as people are forming those

relationships on the internet, oftentimes that is the person that they look to when they want to develop closer connections or feel like they're being heard. And you see a lot of this type of online relationships replacing the kinds of offline connections that we used to have.

Carl Miller: If I may, I mean, we are seeing people fall in love with large language models, when they know they're large language models. There are entire spaces on Reddit dedicated to people using GPT for therapeutic purposes. We are very happy and capable of actually developing strangely deep and meaningful connections with things we know are not human. And this is just this strange emerging trait or proclivity that's come out of this. So yeah, I don't think there's anything that we can see in the way in which we're engaging with these models, which implies that you could not use them to create great shifts which are very meaningful to the people that are part of them. Even if they suspect. Maybe there's a strange behavior. I think in many cases I think people will bury it.

I was going to add, I think that is the reason why we badly need to move away from disinformation as being the idea that's coordinating all these efforts. It's a horrible way of describing the problem. The problem is not that there are lies propagating around online. It doesn't describe the campaigns we pull apart. As Renee says, there's so many ways of wreaking all kinds of influence that doesn't involve lying to someone. It's much more about confirming people's beliefs about the world and guiding them in a certain direction than it is ever about telling them something which is untrue to get them to change their mind.

But more than that, absolutely. You don't want a bunch of think tankers defining about what's true and what's not true in the world. And that's not how democratic debate works. The truth is slippery and fiendish and difficult and contested. It's always going to be like that. It always has been like that. So instead, we need to move away from the idea that disinformation is the problem and towards the idea that hidden covert, professionalized and sustained influence operations are the problem. I do not care what British citizens say in the next election online. I care about everything that the SVR or the FSB or any autocratic military or intelligence bureaucracy says in any election, in any democracy around the world. It's those voices and those actors that we need to deny access to our information environments.

And that is really nothing to do with disinformation. It is got to do with who they are. It's got to do with attribution. It's got to do with identification and exposure. That is the new frontline of this. The new frontline is how on earth can we secure information environments to allow us to reveal when there is sustained and concerted attempts to try and manipulate them by sophisticated bad actors that have absolutely no interest in the health of those environments. And to me, there's two ways forward there. One, online researchers, people like me, people like Renee, and I'm sure many people listen to this. We have to join up more towards investigative journalism. There's only so much we can reveal with all of

our models and all of our patterns, and I could geek out for hours and talk about semantic mapping and how that's changing the way in which we detect campaigns today. And it is. We're getting much better at detecting these things as well as they're getting better at doing them.

But none of that's really going to matter very much unless we can actually uncover the organizational financial realities behind the information maneuver. And that requires journalists knocking on doors, calling people up, forensic accounting, actually doing all the things that journalism is doing. And it does require I think probably the interaction of states and platforms. States have to do more to reveal who is behind these accounts. And they probably have to require more information from people in the first place when they're setting these accounts up. There's much more that needs to happen in order to, especially in the context of elections, to allow us to trace these things back when they should be traced back to bad actors around the world. So that's the new coalface in my opinion. Which I hope is bipartisan, totally uncontroversial position.

Aza Raskin: I think in the Facebook files, when Frances Haugen did her whistleblowing, it turned out that Facebook had found that there was one thing that they could do that would do more to fight all of the worst content, hate speech, misinformation, whatever, than the tens of billions of dollars that they were spending, and it was take away the reshare button after a piece of content had already been shared once. So I share it to you, you get a reshare button, you click reshare, it goes to another person, they don't get a reshare button. That one thing was the most effective intervention that they found because that which is viral is more likely to be a virus.

Carl Miller: Definitely. And platform mechanics for sure, but also I would say around account join up and sign up. There's obviously another clear series of incentives there around making that just as frictionless as any platform possibly can. And actually I think friction is great. I'd like to see a lot more friction in terms of accounts joining up. I'd like to see more challenges and possibly even in the immediate run-up to elections, a really slowing down of who can join any information environment and who can intervene in those kinds of discussions.

Aza Raskin: I wanted to just return for one second to the arc you were telling about what tools were available and have ceased to become available, like the Twitter fire hose. And I want to give space for both of you because we have a lot of policymakers on the podcast. We have a lot of people inside of the companies. I wanted to give space to both of you to say, "What do you need?" In order for this election to go well, please make a very clear and direct ask of what you need to do your work the best.

Carl Miller: The ask is absolutely clear. We need the data and like time and time again, platform after platform, we are losing it. It's as simple as that. It's either

becoming more expensive, it's becoming completely unreachable, it's becoming impossible to deploy advanced analytics on. It is a whole series of different barriers. But that is absolutely without doubt in my mind, the biggest change in the environment in terms of us as the defensive side between 2020 and now. And we can see countervailing forces ahead. There's the Digital Services Act in the EU. There's the UK Online Safety Act has come in now. There's various new regulatory structures which will require platforms to make data available for exactly this purpose.

The open question though is whether any of that will come into force for next year and in time. And my fear is that it won't necessarily. It is hard to overstate how reliant policymakers and regulators now are on this whole ecosystem of researchers that have grown up around trying to detect this stuff. It's so important. Thousands and thousands different ways that research drives the decisions, the concerns which are being raised, the ways in which communities are reached out to the different responses which are being explored. And when data goes away, you can't see it. But that whole ecosystem just goes blind and the research dries up and the decisions become less informed and less evidence-based. That is beyond measure I think my greatest concern for the next year.

Tristan Harris: Carl, just to make sure we're putting an underline on this, I believe you had mentioned earlier that the Digital Services Act doesn't come into effect until 2025. Is that right?

Carl Miller: The Digital Services Act is in effect, it's just a very slow ratcheting up of regulatory action. I think regulators in general take a while to get going. So yes, I think there's a real race, and my big fear really is I think it's quite likely at this point that we are not going to see the level of regulatory action over the next year that we need to defend these elections specifically.

Tristan Harris: I think the line that you had used was that these are the most vulnerable elections in history because even if we have the right protections potentially, and they're not even fully protective, we're in this gap where they're not going to be enacted regulatorily in enough time to actually affect all those elections in time. So we're in this window where we're unprotected.

Carl Miller: We've got a regulatory gap. We've got a whole series of platforms that have fired teams, shut down APIs, pivoted away from their commitment to responding to one of harms in general and protect elections specifically at the same time as this year or so when regulators get up to steam, it is a gap. The next year is a gap where we might actually see perversely less activity actually being done than last time around, even though obviously the trade crafts and the offensive measures become way more sophisticated. Everyone's had so much more practice.

Aza Raskin: Just a heads-up. In a second, we'll be hearing from Renee about Community Notes on Twitter or X. What you need to know is that Twitter introduced Community Notes as an alternative to having content moderation teams. The idea is to crowdsource reactions to tweets to determine what is true, or at the very least, what is agreed upon truth. How it works is that the algorithm gives a higher ranking to comments with greater consensus from users who don't normally agree. And sometimes the Community Notes have published corrections to tweets even from Elon himself, but it's a long way from being reliable. In fact, a recent investigation by WIRED found that it is a targeted self to coordinated manipulation.

Renee DiResta: I think the thing that I'm most concerned about is the way in which some of what we've seen in the Israel-Gaza conflict, I think is illustrative of what's going to happen in the election. And by that I mean you have massive gluts of content that are processed by primarily influencers. They don't actually necessarily know what they're talking about. They're possibly not in region. They just take a clip from Telegram and repost it somewhere else and make it go viral. Often the context is wrong. There's a belief that something like Community Notes will solve these problems. But again, the rumor is going to go viral before the correction appears, even if it's not a correction from a journalist or a fact checking organization, which are inherently distrusted by half the American population, even if it is provided as a correction through Community Notes.

Community Notes can actually tell you if something real happened in place A, B, or C, that's just not the model that it operates under. It's great for adding context to things that are known or for correcting the record or correcting a misleading claim after it's aged for a while. But there is this problem of that gap between when the rumor goes viral and the truth can be known. So I think finding ways to enable counter speakers, enable people who do know what is happening to address it as quickly as possible is really important. But then the other thing that we're seeing, re: the Israel-Gaza conflict is the discrediting of real information as fake.

So this is the flip side, the so-called liar's dividend to generative AI, which is that if you don't like something, you can simply declare it to be an AI generated fake, and then you have absolved yourself of having to believe it. And that is something that we have seen with some pretty horrific atrocity footage in the context of the Israel-Gaza conflict. The willingness of people to simply dismiss something because they can reconfirm their priors or feel good about themselves as being on the righteous side by discrediting a real image is something that is actually horrifying.

So right now, we're still in a stage where a lot of the AI generated content is somewhat detectable. That's not always going to be the case. But right now, most of the AI generated content that has gone viral has had tells and is relatively quickly uncovered. But the flip side of that ability to discredit actual

reality, it's a crisis of trust and a crisis of social divides and bespoke realities. And that is a problem that is exacerbated by technology, but at this point, leaders and influencers within particular communities are actually profiting from it. So their incentives are actually to keep that division going as well.

And I think that again is more of so much of the processing of what happened with Russia in 2015 to 2018 was in the context of the US election because that happened in between. But the overwhelming majority of the content was not political, and it wasn't focused on Donald Trump or Hillary Clinton. It was focused on the idea that you could create deep, strong identity-based communities, reinforce pride in those identities and then pit those identities against each other. And that model has proven to be, I think, quite effective. State actors are only exacerbating things that we have already done to ourselves and that our domestic political conversation continues to reinforce here in the United States. And that is in very effective as a vector for anyone who wants to both obtain profit power or clout by engaging on social media. So not a social media problem, but exacerbated by social media in a bad feedback loop.

Aza Raskin: What are your biggest fears? What do you think the biggest threats to our election are in this next, 11 month time window?

Carl Miller: My fear is bad actors will weaponize relationships, build new workflows to reach target audiences that will answer people's sense of alienation, that will speak to people's swirling sense of loneliness and being by themselves. They'll use those friendships to recontextualize people's grievances, make them feel like they are part of a wider struggle that has something to do with their identity, and use their sense of that feeling of struggle to drag a lot of people into these parallel epistemic worlds. Ones that have nothing to do with journalists that have nothing to do with academics, professional politicians or anyone else. And in those epistemic worlds, ones which are conspiratorial and ones which are radically rejecting of the main way in which we verify knowledge and tend to manage public life, it will make people feel that the elections don't matter. There's no point participating and they were rigged anyway. So it will just de-legitimize the entire process.

I'm much more worried about that actually than simply flipping a vote from one candidate to another. I think it's much more this rot at the very fundamental idea that these elections matter, that they're not foregone conclusions and that they were fought fairly. I think that's likely something that's likely to be the playbook. Let's see. I don't know if we'll be able to detect it, but that might be something we see rolled out again and again over the next 12 months. And not something just for UK of course, lots of these actors will just pick up their suitcases of influence and go onto the next election afterwards, just like any other political coordinator or campaigner. And if we go around chasing these deepfakes and trying to push out the lies, we will be completely

misunderstanding, in my opinion, the models and the ideas of information warfare which will be ranged against us.

The last thing I would just say to the policymakers trying to imagine the threats, is to think like an attacker. That means think imaginatively. Think about how you can try techniques that you never have before and think about all the different vectors of influence which are at your disposal. Why are people not looking at Wikipedia? There are so many different, extremely vulnerable, extremely central and important information environments that we tend not really to look at at all. Especially when you imagine that you can tie in Wikipedia with actual front organizations and cyber offensive measures and buying local media outlets and bribing some influencers. These are the kinds of attack options, which I'm sure sophisticated actors will be laying out as a portfolio for their campaigns.

We must not think like defenders because we as defenders, we will research our platforms. We have our actually quite narrow furrows often of experience where we'll try and detect this, but actually we are way less good really at trying to understand influence across all the vectors in which it can be actually conducted than I think that people are actually doing it. So imagine, everyone has to imagine. They have to use imagination going into the next year. We must not allow our idea of the new threats coming down the line to be defined by the ones that we've already seen.

Tristan Harris: If we were to do the maximum positive things that we could do with all of this, just imagine the comprehensive suite of interventions with the 80-20 rule of what's the 20% of work that we could do that would lead to the 80% benefit, maximum benefit in the face of this to, again, not just hold up our shields against disinformation, but to design for trustfulness, designing for synthesis of communication for bridge rank rather than in personalized engagement, moral outrage rank. What's the full suite in your view of solutions that would bring us closer to that? Where I think of instead of just going on the defense and holding up shields to an infinite tidal wave of new threats, but instead actually asking what's the offensive set of comprehensive assertive measures that we could do in your view?

Carl Miller: Well, I've said my piece around the data. But that's an absolute given and I think we can't repeat that enough really. There needs to be access to basic ways of being able to spot when bad things are happening on the information environments that matter. But I think to me, apart from that, the solutions actually lie in asymmetric non-information responses. I think we have over-focused on trying to upskill hundreds of millions of people to try and spot this with digital literacy. I don't think we have enough time. I don't think we can reach the people that we need to, and I don't think people could spot it even if they're taught to. And I don't think what they can spot today will be spotable tomorrow.

And I also think simply knowing that these operations exist doesn't work, because it's actually when they are confirming your worldview and flattering your idea about the world, that's really when they're working and that's not when we're on our guard. To me, there needs to be, and this is going to sound strange for someone that comes from a center left think tank and spent 10 years in there, we need more activities from states to levy more costs and risks against the specific professional bad actors doing this. We need to put people on no-fly lists. We need to sanction people. We need to look at criminal laws and we need to degrade assets. We need to make it harder for these campaigns to access Western finance. We should deny them the whole tech stack. We need to deny them app stores. We need to deny them operating systems. We need to deny them search. We need to squeeze off their audiences.

And I don't really think that primarily means we need to be maneuvering in the information space alongside them. I actually think we need to grow this whole other portfolio of responses. And some of that's going to be states, sometimes think tanks, sometimes law enforcement agencies. I think there's a whole mixture. But that can basically over time just make these operations less effective and less profitable and less easy to do and riskier to do. That's the only way that we can begin to swing the strategic balance in our favor. Because at the moment, what influence operator has had any serious repercussion from doing what they're doing?

We worry about this as being one of the most formidable threats to our democracies that we're currently tangling, and yet we have not managed to really levy any serious costs against any of the people that do it. And to me, that is a mad imbalance. We as democracies levy cost against all kinds of people that I think do all kinds of things that aren't as dangerous, I think as we feel some of these threats are. And we've got to change that. So that to me is actually exploring a whole series of non-informational responses.

Renee DiResta: I've been spending a lot of time in these old archives from the 1980s to the 1930s, the Institute for Propaganda Analysis and the Active Measures Working Group. I have this book coming out in June. And I went through these old archives because I was curious about this question about what do you do about it? Because again, the medium is different, the means is different, the extent to which it's much more personal is different. But what are the ways in which we've looked at this in the past? And one of the things that I appreciated was the way that, so in the 1980s Active Measures Working Group exposed Soviet influence operations. It was started by Reagan and Gingrich. So there was no partisan valence to it actually, the right was at the vanguard of this.

And what you saw there was the US Government transparently releasing all of the information related to an operation to reinforce to the public that it was happening. Here's what they did, here's how they did it, here's how it was executed, here's who picked it up. And this is a very interesting model because it

also was done at a time when trust was higher. So let me put a pin in that for a second. Then you have back in the 1930s, the Institute for Propaganda Analysis, which is a civil society effort, and that's a bunch of professors from up at Columbia and elsewhere. And actually they were concerned about the rise of domestic fascism in the United States through influencers like Father Coughlin.

So a lot of what they did, and I absolutely love these documents. What they do is they write a guideline for recognizing the rhetoric of propaganda. Here is how when somebody says the word, they. They are probably manipulating you. Here are the weasel words. Here are the signals. It doesn't matter who it comes from, whether it's Father Coughlin or some other demagogue or somebody in Germany for that matter. Here is how you need to think about these signifiers, these words. So the explanation isn't media literacy around here's how to use the internet or here's how to detect a GAN generated face because as Carl notes, these things evolve quite rapidly and pretty soon they're not going to be very easy human detectability.

So the question then becomes like, how do you deal with what emotionally resonates about it? And I think that those are the two areas that in recognition of the fact the technology will keep evolving and keep changing, and again, we can watermark until the cows come home. It's not going to solve the problem. It's important. It's useful. Won't solve the problem. So the question is how do you address that crisis of trust? And can you do that through these sorts of transparent programs that aim to explain how the rhetoric works? Why does this make you feel a certain way? Why does this make you angry? Why does this make you feel good about yourself? And is there something that you should be paying attention to with, for example, excessive flattery?

So I think that that set of lessons is critical and actually a surprising lost art I think. These pamphlets that the IPA produced were given out to middle and high school students. They were shared at the local bridge club. This was the thing that was just considered like a patriotic education in rhetoric. And then ultimately it was shut down as the US entered World War II. And the guys who started it got caught up in the Red Scare. They were investigated by Congress. So remarkable parallels to the current moment.

Tristan Harris: I want to thank you both for the incredible work that you do and raising awareness about these topics. And there's obviously so many more things to cover, but thank you so much for spending the time and educating listeners, and I hope policymakers hear you and take your advice.

Carl Miller: Thanks very much.

Renee DiResta: Thanks for having us on.

Aza Raskin: I want to wrap up by underlying the precarity of the situation. We're at our most vulnerable, with less protections than we had even in 2020 while AI makes the threats the greatest we've ever had. But it isn't hopeless. There are some really clear steps that we can take right now.

Tristan Harris: Starting with platforms, Twitter must reopen its research feed, to all academic researchers. We need to demand more transparency as you just heard in this discussion. After Elon came in, he changed the policy so that researchers have to pay $40,000 a month to access the Twitter feed, and they're limited queries. So it's almost impossible to know what's really going on at scale. And Facebook needs to open up its ads API, so that all ads are available for public scrutiny to journalists and researchers under the Facebook API, not just for political ads.

Aza Raskin: We talked to Frances Haugen, the Facebook whistleblower, and she gave a number of recommendations for other things that the platforms could share. Their levels of staffing or sharing their operational metrics. For example, what fraction of attackers are ever taken off, come back again as recidivists? What fraction of influence operations are identified by external versus internal reports and what fraction of the threat are actually being taken down?

Tristan Harris: And just like we have media blackouts in some countries before a major election, we could have digital virality blackouts where we don't make things go viral indiscriminately for certain periods that are sensitive and more delicate. So whatever process we're talking about should be bipartisan oversight, have public transparency, and it should be in the name and good faith effort of reducing this engagement monster that creates basically an unwinnable game.

Aza Raskin: What is the opposite of engagement? Well, it's latency. It's like putting in a pause. You can't continue engaging. So if we want to hit the engagement companies where it hurts, it has to hit them in engagement.

Tristan Harris: For platforms that have experimental or uncontrolled features, like during the 2020 election, Facebook promoted live video because those teams were getting incentives for driving more engagement on the platform. But during sensitive periods like elections where you have features where you don't really know how that live video is going to affect things, platforms could turn down the engagement on those more uncontrolled untested features that don't have a lot of verification about how they perform in these sensitive environments.

Aza Raskin: There's an interesting solution direction that comes from the US stock market, which is when you get these flash crashes, when the market just starts losing a whole bunch of its value, there's literally a circuit breaker. They just pull it and it pauses trading for 15 minutes, 30 minutes the rest of the day. You could imagine something similar in social media where when you get near these sensitive times like elections, you just switch from an engagement based ranking of your feeds to a chronological feed. You just tone down the engagement.

Tristan Harris: Thank you so much for showing up for this podcast in 2023. There is a lot that has to happen in the field of AI for the world to be shifted to a different path that leads to a better future. And we have a lot of exciting things to share with you in the new year. So we'll catch up with you then. *Your Undivided Attention* is produced by the Center for Humane Technology, a nonprofit working to catalyze a humane future. Our senior producer is Julia Scott. Kirsten McMurray and Sara McCrea are our associate producers. Sasha Fegan is our executive producer. Mixing on this episode by Jeff Sudekin. Original music and sound design by Ryan and Hays Holladay. And a special thanks to the whole Center for Humane Technology team for making this podcast possible.

You can find show notes, transcripts, and much more at humanetech.com. If you liked the podcast, we'd be grateful if you could rate it on Apple Podcast, because it helps other people find the show. And if you made it all the way here, let me give one more thank you to you for giving us your undivided attention.