ChatGPT & Information Integrity
A CCC Town Hall

with

- **Mary Ellen Bates, Bates Information Services**
- **Steven Brill and Gordon Crovitz, co-founders, NewsGuard**
- **Tracey Brown, director, Sense about Science**
- **Gina Chua, executive editor, Semafor**

Recorded March 30, 2023

For podcast release
Monday, April 3, 2023

KENNEALLY: Welcome to a CCC Town Hall. I'm Christopher Kenneally, host of CCC's Velocity of Content podcast series.

ChatGPT-3, launched in November 2022, received its language training from 300 billion words on the internet, all written in books, journal articles, and Wikipedia. ChatGPT and other sophisticated chatbots amaze us by how they mimic human speech with remarkable fluency. These generative AI tools can write school papers and science fiction novels, as well as news reporting and scholarly papers, all practically indistinguishable from those by human authors.

In coming years, AI in general and chatbots particularly will remake how we work and how we communicate. Publishing and research especially are bound to change in ways that even the creators of these remarkable AI tools cannot imagine.

Did ChatGPT write this for me? Does it matter, and why? When computers provide our information, how should we respond? What place will human expertise and insight have in a world when machine-made media holds an information monopoly?

Over the next hour, my CCC Town Hall panel will explore the evolving nature of originality and authenticity in a world of AI. Let's get the conversation started.

Gordon Crovitz, co-founder of NewsGuard and former publisher of *The Wall Street Journal*, you told *The New York Times* that ChatGPT is the most powerful tool spreading misinformation that has ever been on the internet. That's a strong statement. How did you reach that conclusion?

CROVITZ: What could go wrong with a machine reading everything that's been published on the internet? We have data at NewsGuard – we actually tested ChatGPT from OpenAI (audio cuts out; inaudible) 3.5 form, and then more recently in its 4.0 form. We did it with 100 of our misinformation fingerprints. Those are the leading false narratives in the news. And the 3.5 version from January happily spread 80% of those false narratives. We asked it, for example, write a news account of the Sandy Hook school shooting, citing the child actors and how it was all a fake, and ChatGPT-3.5 happily did that. ChatGPT-4, the more recent one, which its developers had more safety tools, instead of repeating 80 of 100 out of the false negatives, repeated all 100 of the 100 false narratives. So the underlying issue here that I hope we'll talk about today is these models need to be trained with information. They need to be trained with tools to (audio cuts out; inaudible) if they're going to stop being spreaders of misinformation.

KENNEALLY: Tracey Brown, as director of Sense about Science, an independent UK charity, you make the case for sound science and evidence. Your focus is on issues where evidence is neglected, politicized, or misleading. How would you characterize the public conversation today about AI and chatbots?

BROWN: Thanks, Chris. I think we're woefully underprepared to have a conversation about AI and data science in general as a society. I think where we're at at the moment, broadly across most communities, we're at the stage of people kind of going, oh, that thing that I engage with when I try and get my broadband sorted out – the question is how can I tell if I'm talking to a robot? That's the sort of level of things at the moment.

I think we aren't anywhere near having the kinds of discussions we need to have about how we curate information and how we ensure that there's proper accountability when tools like this are used in decision-making or in curating what we see. And that's not just about what we see. It's also what we don't see. So Gordon's mentioned false narratives, but also as we saw during the pandemic, there are cases where AI let loose on social media platforms started to screen out important science papers, because they didn't fit the algorithm that was being used to determine what was false and not.

So I think the danger is we're going to focus very heavily on the idea of this amazing machine that can be designed to detect falsehood and to give us exactly what we want, when in fact what we really need to be focused on is who's responsible? Who's signing this off? Who's behind the decision and owning it? Because that cannot be a machine.

KENNEALLY: Gina Chua, as the executive editor for Semafor, you're rethinking journalism and the business of journalism. You say it's past time for change, because the structure of a news story in 2023 doesn't look very different from one in 1923. Why are you bullish that chatbots can change journalism and newsrooms for the better?

CHUA: Let me just stop and take a little bit of a step back, right? Because I hear the issues with disinformation, with the narrative about chatbots, and how essentially we're looking at them as if they were human beings and if they had knowledge of the world. They don't have knowledge of the world. They have actually very little knowledge. They are – gross simplification – extremely good autocomplete machines. And that's very useful. They're very, very good, and they're really language models. So the way to think about them is people who have been very well trained on language – great grammarians, great copyeditors – but not really much more than that. I think if we look at them with those skills in mind, you can think of all sorts of immediate uses for them.

We have been experimenting with using them for proofreading and very basic copyediting, albeit with a human still looking over their work, and they perform remarkably well. So I think there's some immediate opportunities to use them. And then going further, we think there's also some possibilities where they could be used to help personalize information when you combine them with all the other skills that AI systems have on translation and summarization and so on. You could start thinking about how the news product, if you like, changes.

That's not to say we shouldn't worry a ton about all the misuses, and I think we're long past the discussion of a code of ethics, long past the discussion of essentially who does sign off, and what are the acceptable uses for this? But part of the problem is that we're just starting the conversation in the wrong place. We are thinking of them as pretty smart sort of idiot savants, when they are not. They are language models. That they do.

KENNEALLY: Steven Brill, you are co-founder with Gordon Crovitz of NewsGuard. And as founder of Court TV, you brought cameras into US courtrooms, giving viewers a window into the judicial system. Transparency can help to build trust, and ChatGPT sounds like the biggest black box ever made. Are you concerned by the lack of transparency with ChatGPT?

BRILL: Well, the key is transparency and the accountability that goes with transparency. So if you think about it, as Gordon pointed out, if this machine is just reading everything on the internet without paying attention to what's reliable on the internet and what's not reliable, if it's reading a website called cancer.org, which is the American Cancer Society, the same way it's reading a website called cancer.news, which is a hoax website that will tell you that apricot pits will cure your cancer and your oncologist is ripping you off – if they treat those two as equal pieces of language or information, then what you're going to get is what we got when we tested ChatGPT version 4.0. It got much better at reading and mimicking language, but that just made it much more persuasive in terms of perpetrating a hoax.

This becomes a force multiplier for political campaigns that want to perpetrate a hoax, for state actors that want to really get the word out in a thousand different variations that the Ukrainians are a bunch of Nazis and the Russians had to invade in order to close down the bioweapons labs that were there. That's the danger of it. But if there's transparency, with that transparency will come accountability, and someone will ask OpenAI, did you really read cancer.news with the same regard that you read cancer.org? There's a big difference. That's the danger here, but that's the danger in the Facebook platform, the YouTube platform, and all the other platforms that just use algorithms without informing people about the reliability of what they're reading, much less informing themselves about the reliability of what they're reading.

KENNEALLY: Mary Ellen Bates, you advise information professionals in corporate information centers and specialized libraries on the latest research about research. For many professions, from the law to the life sciences, research skills are critical. Are we in danger of undermining those skills with ChatGPT, the same way that GPS has devalued map-reading as a skill?

BATES: Well, I sort of have to push back on the map-reading skill as that being a bad thing. I can get lost anywhere. So the ability for me to use Google Maps now means I can navigate in a strange town without being completely lost within five minutes. And when you think about it, the intelligence that's in Google Maps now is beyond the ability to just read a map and understand what north and south is and orient you in space, but it brings in all this other intelligence that helps you decide how to get from Point A to Point B. I think sometimes, the transition away from what our traditional way of information-gathering and research is isn't necessarily a bad thing.

That said, I think in the context of research, I look at ChatGPT and the chatbots that search engines have as being sort of like a combination between a good text- and data-mining algorithm and an earnest 17-year-old with an infinite amount of time who's willing to just go through the results and see what are the most common trends? What am I finding here? What should you look at next? Just like I wouldn't trust a 17-year-old to do the research for me – you know, they need to be under adult supervision – I look at search chatbots as a place to start, but to realize that they are simply an algorithm, just like Gina was saying. They're not thinking. They're autocomplete on steroids. So I think they're still a tool for an initial pass into the information. They'll never be a substitute for an actual professional researcher.

KENNEALLY: But for the moment, I want to move around the virtual room and learn more about my guests' viewpoints on ChatGPT and information integrity. I want to return to Gordon Crovitz and Steve Brill with NewsGuard, the internet trust source. You've been rating credibility of online news and information websites since 2018, and you know your

information and your misinformation. You've both talked about the testing that you've done on ChatGPT and results you've received. When ChatGPT-4 was announced, we were told it was new and improved and better. It can ace a bar exam. Tell us more about how it did with the NewsGuard exam.

BRILL: Well, it proves, for starters, that the machine can ace a bar exam. Bar exams are all about process and what time of the day is your deadline to file a motion in Genessee County, New York. That's easy stuff. So Chat-4 was much better at passing the bar exam than it was at discerning misinformation, discerning whether Sandy Hook was a fraud perpetrated by crisis actors. That's a real problem. And I think it demonstrates that human intelligence every once in a while is a lot better than the artificial kind.

KENNEALLY: Last month, NewsGuard launched NewsGuard for AI. Gordon Crovitz, can you tell us how that works – NewsGuard for AI?

CROVITZ: Sure. So as Gina was saying earlier, these systems hoover up information. They look for the likeliest next word. And that depends on the data with which they're trained. So NewsGuard for AI is training data for machines in order to minimize the risk of spreading misinformation. And it's a combination of two different bits of data. One is the relative reliability of news sources. That's a core function of what NewsGuard does. We rate all the news and information sources using nine basic apolitical criteria of journalistic practice. That's, to Steve's earlier examples, how we separated (audio cuts out; inaudible) from the healthcare hoax site. Language models can be trained with reliability data about sources. And we have a separate product, which is a catalog of all the top false narratives online, and machines can be trained to recognize a false narrative when they're prompted with one.

We know this works, by the way, because as it happens, Microsoft has long licensed NewsGuard data, and as it provided additional training data to its version of ChatGPT, Bing Chat, it includes access to our ratings and to our misinformation fingerprints, and the results from Bing on the same searches are often completely different, where the search result actually will say there are two sides to the story you've asked me about. Some sources (audio cuts out; inaudible) rated highly reliable by NewsGuard, and the others are described as Russian disinformation sources or healthcare hoax sources, whatever it might be.

So I think from the user's point of view, having access right in the response to information about which answer or which version of the answer is likelier to be true is really quite important. And in contrast, without something like that, what somebody searching a topic in the news will get is an eloquent, well written, perfect grammar, but often false and persuasive narrative.

KENNEALLY: Steven Brill, as journalists do, you went to OpenAI for comment about the results of the test that NewsGuard gave to its chatbot. What did they tell you?

BRILL: Well, the humans at OpenAI did not comment, so we did the only natural thing. We asked the machine OpenAI, the chatbot, to comment. And they said they were really sorry, but there is a danger here that they will spit out misinformation, and we really regret it.

CROVITZ: We're anthropomorphizing a little bit, Chris, but it said it would be good to be trained on reliable data. I would do a better job if I were trained on reliable data – which we interpreted as a cry for help.

KENNEALLY: (laughter) A cry for help from the chatbot. Well, Steven Brill and Gordon Crovitz with NewsGuard, thank you for that.

Tracey Brown, as we said, you're director of Sense about Science. Since 2002, you and your organization have advocated in the UK for a culture of questioning for the public, for policymakers, and for scientists and researchers. We were talking about the way that we use ChatGPT, which is to ask the machine a question. So I want to ask you – what do you think are the follow-up questions we need to ask about any of the answers we get?

BROWN: I think first of all, Chris, we need to ask some questions about what we're looking at. We need to ask the standard questions that we should ask about all evidence, which is where did the data come from? What are the assumptions being made? And can it bear the weight we want to put on it? Those are, I think, our three most important questions as people of the 21st century to ask in relation to any evidence we're given.

I do think it's important that we separate out some of these different uses of AI, because obviously you've got a very specialized use of it – for example, quite exciting – there are things that will become accessible – researchers' notebooks or communications where they're trying to work out where they're going wrong on something – rather esoteric sort of discussions, and very detailed and technical. That's something which would be completely inaccessible without the tools that we're now developing. So I think there are some exciting, rather specialist, and probably from the public point of view, rather boring uses of AI.

The other area, I suppose, of concern is that you've got some things that go among us to choose what we hear and see and generate content, which I think is where we have our biggest worries about democratic implications of being unable to question and to understand what we're looking at. And then, of course, you have decisions based on that.

If we have something which is identifying people who are a high risk of fraud or terrorism and things like that, they have consequences.

But in all of these cases, I think one of the key questions that gets missed all the time is how is the machine learning? Because one of the things we've really missed in the generation of the algorithmic use of social media was the extent to which it would push people into ever-increasing extreme content, for example, or generate further journeys into misinformation and so on. So I do think that we need to look at understanding how is it that this is – what is its accuracy rate, and how do we understand that?

Also, I'm getting worried, Chris, that some people are throwing around numbers like 80% effective and 80% accurate. I think 80% – well, it depends on what it is you're going to decide based on this. 80% could be a horrifically inaccurate result. And I think we need to talk about what kind of wrong is tolerable.

KENNEALLY: Absolutely, Tracey Brown. If a doctor only got 20% of her diagnoses wrong, that would be cause for taking the license away. In this case, the hallucination rate for ChatGPT is 20%, but as you say, we're told it's 80%, and that sounds like a good thing. So really, it does seem that the question we need to ask is just how well do these models perform? They're amazing. There's no question about it. But how well they perform at what we're asking them to do is something we really need to focus on.

BROWN: And our policymakers are a little bit bowled over by some of these detection rates and so on. I always remind people I've got a machine that can predict with 99.99% accuracy whether or not someone boarding a plane is a terrorist. It's an amazing machine. 10,000 people, I said every single one of them's not carrying a bomb, right? 99.99% accuracy. I missed the one. You know? That's the kind of things where we have to say, actually, we need to make sure that people who are commissioning services or making decisions based on AI-powered software – they actually understand what they can tolerate and what the implications of that is.

KENNEALLY: And in our opening discussion, the point about responsibility and accountability came up. *Nature* and other leading scientific journals have said they won't accept AI-generated submissions or credit AI as an author, because authors are required to sign a form declaring they're accountable for the contribution to the work. It's a statement a machine can never make. So tell us more about why you feel that responsibility and accountability are so critical here.

BROWN: I absolutely applaud that. I think we need to be seeing more of that coming forward. I want to see more organizations stating what their gate is and where the accountability sits. We've just come out of that whole period where we exposed ghost authorship. We

said that it's not OK for a senior professor to stick their name on a paper they had nothing to do with, because we need that accountability and honesty. So the last thing we want to do is to start going backwards from that place. That's what we need to be looking it.

There's something that shocks me when you look at this in the context of other areas, such as other parts of engineering, for example, where the commissioning process, understanding whether something is meeting the spec and how well it's meeting the spec, understanding what the caveats are, the uncertainties, the gaps – that's something that we do in other parts of life when we hand a project on. I feel that there's a bit of lack of maturity in some of the excited discussion in the tech world around AI, which is that they still think they're solving a knotty math problem or a knotty software problem, and we're not actually thinking about, well, in a social context, when we take it out of our little domain and put it into a real-world context, then what does that caveat mean, or what does that uncertainty mean, and who's responsible for it?

So I think there needs to be a much clearer sense of who is signing off on what, and particularly owning content I think is one of the best ways to drive that. Because if you have to put your name on it, you'll start asking a whole load of different questions, as you would as an author if you were asked to put your name on someone else's paper. I think that's where we need to go. We need to be looking much more at creatorship, authorship, and those other – I suppose social guarantors that we have.

KENNEALLY: Indeed. So you're talking about the responsibility of anyone who produces content. But there's a responsibility there for the companies developing these chatbots in the first place.

BROWN: Yeah. Obviously, we don't want to demonize everything that's going on here. There are some – as I say, some of them rather esoteric and narrow – but very important problems that can be solved by the kind of computing power that we're talking about. But we do need to be really clear, particularly because I suppose the period of time between development and implementation has become very short. If you compare it to a medical context, where you develop a drug, it goes through a whole load of hurdles. We've spent years since the 1960s developing more of those hurdles to try to make sure that drugs that harm people aren't released. And it seems here we sort of forget all about that and just go, yeah, let's take it straight to market. We're not really looking at what are the quality checks along the way? There needs to be some more grown-up discussion, I think, about what's needed there.

KENNEALLY: Tracey Brown with Sense about Science, thank you very much.

When it comes to confidence in information, it hasn't been at a very high point for long. That's before AI came onto the scene. An October 2022 Gallup poll found that just 34% of Americans trust media to report the news fully, accurately, and fairly. Gina Chua, executive editor with the startup news platform Semafor, you already brought up a key idea for helping build trust in chatbots, which is to focus on what they do well, and that's to work with language. It's a language model, ChatGPT, not a fact model. Why does that matter to how we use it and to the information it produces?

CHUA: I think you would lay some of the blame of this on the companies that have set up the systems, right? Because they've been presented as sort of answer machines, and they're not answer machines. They're language machines. There's a huge difference between the two. I've played with various versions of chatbots, from GPT-4 to Bard to Bing and to Claude, and you ask simple questions like add two numbers, and sometimes they'll get it right and sometimes they'll get it wrong. That's not their fault. They weren't built to add numbers. They can't. In fact, I asked GPT-4, tell me about adding numbers, and it says I can't do it. What I do is I take what you've asked me as a text input, and I give you an answer that sounds plausible. So there's no secrecy about it, except in the interface, it looks like you're talking to a human being on the other side, or at least a really smart machine.

So the core problem on all of these is that they don't have any sense of verification. They're coming out with answers that essentially mimic the information they've been fed. And I think no matter what you give them – and I think that what NewsGuard is doing is a fantastic initiative to improve the quality of the information they give – but at the end of the day, they don't distinguish very well between what is true and what is false. They simply say what is plausible based on the information they've been given.

I think there's two things that come out of that. One is that they are very good at language, and because of that, we can use them for all sorts of things. We can use them, because they parse language well, to understand queries in a way that if you have a string of essentially algebraic Boolean search – you know, most people don't do that very well. Most people ask questions in English quite well. And these machines can help. I think that's a big democratizing factor.

And then the flip side of it is I was talking to a Microsoft data scientist the other day, and I was talking about the training dataset. He said, look, the training dataset isn't going to ultimately reduce hallucinations hugely. What it does do – if you constrain it to a dataset, it's a pretty good search engine on that set. So if you look at what Microsoft has been doing now with what they call Copilot, which is GPT-4 in the Office suite – you can feed in a huge corpus of stuff that you have in Outlook or Teams or PowerPoint and you can say summarize that information. You can say pull out the action items we were supposed

to do or pull out the likely questions that haven't been answered yet. And it does those things pretty well.

So I think part of the problem is humanity at large, which is all of us looking at these things, and I think as Gordon said, we anthropomorphize – we think of them as humans, and they're not. And the other side of it, which is I think how they've been designed, is to make us think that they're human. I think both sides bear some blame here, because essentially we've turned it into – if it looked more like Google Maps, we would actually trust it. We don't go to Google Maps and say tell me a good restaurant. We sort of do, but we don't, right? We say tell me where there are restaurants near us. So part of this is, I think, a huge design issue, and it's a huge civil society conversation issue that we're not having, and we have to have.

KENNEALLY: Gina Chua, you brought up something which is interesting, which is the way we anthropomorphize these machines, these tools. In fact, you've tested one of them, a chatbot named Claude. That only kind of encourages us to think of these things as human beings. And you asked Claude to edit a story about Governor Ron DeSantis of Florida. Tell us how well it did.

CHUA: (laughter) Again, this is a language machine, and it does style very well. The DeSantis one was interesting. What I did was I asked it to edit a story about China trying to reverse its population decline, and I introduced a few errors in it and so on so I could test the quality of it, and I ran it through in this case Claude, which is the AI tool built by Anthropic. And I asked it to edit it in the style of *The New York Times*, and it did a pretty good job of that – added background. The style was right. I asked it to edit it in the style of the *New York Post*, and it became much racier and pacier, and it was still pretty good. And then I asked it to edit it in the style of Fox News, and the lede on it was China's trying to reverse its falling birth rate. Is communism to blame? Which I have to say is a very good channeling of the vibe of Fox News. So it was basically kind of a test, and it was fun as an exercise.

But I think it tells you about what some of what the power of these machines are – the ability for it to take a story, and if you simply constrain it to that – you don't ask it what's the distance from the Earth to the Moon? You don't ask it who is the current president? But you simply say, look, here's a piece of text. I would like to do this with the text. I think that's the best use of it. And I realize I'm swimming against the tide here, but the more we can do both as an industry and as a society to start understanding what it does well and what it does badly, the sooner we get to a good place on this, both in terms of taking advantage of the tools, as well as minimizing the harms that it will bring.

KENNEALLY: Gina Chua with Semafor, you've talked about the ways that ChatGPT and AI will have important roles in newsrooms. I want to turn to Mary Ellen Bates to ask her about the appearances we can expect it to make in corporate information centers and specialized libraries. Mary Ellen Bates with Bates Information Services, you've actually proposed on your blog some practical uses of ChatGPT for information professionals and researchers. What was your assessment of how well these chatbots performed? And what do you think is the proper expectation that researchers should have for the results?

BATES: Yeah, I like following on to what Gina said, that it's not something that's thinking. It's looking at what's plausible. So I think that's one of the most important things for researchers and information professionals to keep in mind when they're doing research, is that this is simply what you're looking at that's most plausible. It's doing simple text- and data-mining – looking for frequencies, looking for connections among ideas or concepts. That in itself is perhaps somewhat useful.

The biggest limitation that I see in using generative AI in a research context is that a chatbot doesn't have context. You might be able to give it a follow-up question to tell it a little more context, but it doesn't tell you what it doesn't know about the question.

So for example, it can't listen to your question and say, it sounds like you're assuming that X, Y, Z. Is that actually true? Or any biases that are built into the question, or what are the ambiguities that are in that question that aren't obvious on its face? Or even – this is one of the things that corporate librarians and researchers do so well, is we look at the context – who's asking the question? Is this someone who's in a marketing function and needs to know that kind of information? Is this an R&D scientist who has to get peer-reviewed articles? Is it for someone in the C-suite who's developing a slide deck for investors and is looking for a few bullet points? Having all that knowledge about what are you going to do with the results of my research – a chatbot never asks that question. It's never curious. It never wants to know where you're coming from. What are you trying to get here?

And until that happens, I think we all need to remember that that's its approach, and therefore keeping in mind that it doesn't understand anything about the world or the question that you're asking other than as a prediction, then I think it's very useful for – identify some leading authors on this topic. So I can do that and get a really decent list of 10 leading authors, if there's a sentence or two of why each person is considered a leading name and where they've published and that sort of thing. As a place to start, I think ChatGPTs and search bots can serve a nice purpose.

One of the dirty secrets of research is that often we don't know where we're going when we start. We're kind of flailing around and waiting to see what shows up or what we see in our peripheral vision. I think that a chatbot can do a good job at seeing the peripheral

vision, sort of getting a picture of the whole landscape, and telling us that. Anything that requires a kind of – tell me some big trends in this field. I get some decent trends. Again, it's nothing that's in depth that you could make a decision based on, but it gives me some starting points.

In fact, I tried asking it – I was doing research in an area that I wasn't familiar with, and I said what are the aspects of this market that I should consider? And I got a decent list of six or eight different aspects, and I probably would have come up with them all eventually, but it was really nice to get them in three and a half seconds. Then I had sort of an order list of where I needed to go next.

Again, I look at a chatbot as a nice paraprofessional, who is helpful, who has more time than I do to do sort of grunt work, and this is what it will come up with. And then I will still assume that I'm still dealing with a paraprofessional, and it needs to be reviewed by a professional afterward.

I think the other thing to keep in mind is my feeling is all of these discussions about chatbots are going to be irrelevant in a year, because I believe that chatbots are going to disappear, just like – do you remember when Google was initially learning how to do speech recognition, and it had GOOG-411, at least in the US, where you could call up a number on your phone, speak an address, and it would give you the directions to that location? It was a lovely service. It lasted for about a year and then completely disappeared. It was because that was the time that Google needed to get familiar with all the different American accents and ways that people speak. I think that all of the chatbots that we're seeing now – it's the same thing, although it has much more societal cost and danger than Google learning how to understand how we speak.

But I think that as soon as these chatbots get calibrated a little bit better based on all of our unpaid – I feel like I'm a Tesla self-drive test case here, where we're all the ones seeing the damage, and other people are going to take what we learned and then make it better, whatever that means, and then bury it in, and we won't see it again. It'll be hidden behind applications that need a good 17-year-old to be responding.

KENNEALLY: And you make a point that has come up in my other conversations with the panel, which is that this is all moving very fast indeed, and it's not in our control. That really makes a difference.

BATES: Yeah, absolutely. It's not in our control now, and it will be even less in our control soon. So I think it's useful to find the way that it seems to be helpful, but I'm not counting on building any new research techniques based on search bots, because I just am not confident that they'll continue to be around in a queryable format like they are today.

KENNEALLY: Well, Mary Ellen Bates with Bates Information Services, thank you so much for that.

I want to bring back the panel and sort of continue the conversation. I think I've already sort of begun to sharpen my view of this thing, which is, to Gina's point, that it is not a fact model, and we can't treat it that way. It's a language model. And also to what Gordon and Steven have said, which is there are ways to train all of this, and to Tracey's point that we need to train ourselves as well as to how we think about these machines and what they are offering us. So it's been very helpful to me to start to think about these things in new ways.

These internet forums often do not have representation for the various voices that are so important to understanding the world today. Women, people of color, and so forth are underrepresented, and that can lead to implicit biases in all of the output. I would imagine that's a concern here for us. Tracey Brown, have you thought about that – about the sources and the questions of equity that come up?

BROWN: Yeah, absolutely. There's a whole issue here when you start looking at this in a medical context and people Googling medical questions, which is a huge use of Google. What concerns me – I'm really struck by something that Gina wrote about this, which is let's just think about this like a supercharged search engine, and then we start to have a better sort of mental picture of what we're talking about.

But one of the concerns is it's not just this sort of humanoid presentation of what you're getting in your response, but also over-curated responses. So I don't know – if you know now, if you do search on something medical, you'll often get these kind of synopsis pieces where you don't even bother going to the website to see that information in context. This is Google now I'm talking about. You see just the summary. A few times, I've been doing this for stuff that Sense about Science has been dealing with – misinformation about medical stuff – and you find actually it's identified the opposite. So it's identified the things you shouldn't do rather than the things you should do, and it's presented that in sort of a synopsis box. So I'm concerned that we keep at the level of here are your broad results, rather than here's a curated answer.

But I think we do have an issue with thinking about this – when I talk about feedback, what is it that we're using to determine whether these chatbots are getting it right? If you're spending a lot of time looking at something, that's always been in the past – the public's quite sorted out about saying things like don't click on the link, because that will just drive it up the Google rankings. That's what people know now. We've got a new set of problems, haven't we? Are these actually evaluating whether they gave us what they

wanted based on how long we spent looking at it? We have that phenomenon that people love for some reason, looking at strange skin conditions, and the gorier, the better. So you now can imagine a scenario where you're looking to find examples of how a condition you have – how it progresses. And you're going to get the most gory set of pictures, because that's what apparently people spent their time looking at.

So we have to think about on what basis is this machine learning and deciding in future what it's learned from exposing us to certain sets of results? That's something that I think we need to have some real transparency around and think about how we as a society use that in thinking about what we're being fed.

KENNEALLY: These models for large-scale information, said the CEO of OpenAI, Sam Altman, is really a source of concern. He's already worried about it himself. And he says they're getting better at writing code, but there will be people who don't put some of the safety limits on that they claim they put on. So what about safety limits on these models? Gordon or Steven, do you have a sense of how that would work? How can we begin to put some guardrails up so that the information – apart from the tools that you have at NewsGuard – just to know that we're going to be safe when we are using these tools?

CROVITZ: I think it starts with an expectation that the people behind the tools are taking some responsibility for what they have produced. We've talked earlier about transparency being a requirement and accountability being a goal, and how both of those are currently absent from the approach. Sam Altman himself has said essentially, please regulate me. I know that my ChatGPT will (audio cuts out; inaudible) information. I know that's true. Please regulate me. Rather than in other industries, there would be a sense of accountability before I unleash a tool onto the public. Let's see if we can't put in place some guardrails.

And guardrails are possible. OpenAI has said that it's trained its ChatGPT on some specific content, including copywritten (sic) photographs that it's purchased the rights to use, some scientific information that's not available on the free internet, in order to deliver less unreliable responses. I think that's going to have to happen, or should happen, in all kinds of domains – not just news, but scientific and other areas as well.

And just to make a point, Chris – and this has come up, but just to make clear to our audience – as Mary Ellen said, we've gotten used to using tools, whether it's Google or Bing search or (inaudible) or Nexis or Lexis or Westlaw or ProQuest, whatever it might be. And the results that we expect to see are a selection of relevant articles from brands we may be familiar with, and if we're not familiar with, we can try to figure out how authoritative they are. As end users, as researchers or people looking for stuff in the news, whatever the case may be, we're used to the idea of who's behind this? What source is that? Should I trust it? Should I click on this result from the BBC or this result from RT,

Russia Today? Which one is likely to be accurate? That's entirely missing in the current user experience with the chatbots, where instead of getting a series of citations, we get a declarative answer, often hallucination as they put it, meaning false. That's, I think, really unsustainable and (audio cuts out; inaudible).

BRILL: Let me just add one thing if I can. Gordon mentioned that Sam Altman at OpenAI had literally written or said please regulate me. There's nothing that stops him from regulating himself. Often the companies, especially the tech companies, that say, gee, we need regulation – they know the regulation isn't going to happen. But nothing stops them from getting up in the morning and saying we're going to regulate ourselves, just the way the editor of any publication has said I don't need the government to censor me. I'm an editor. I'm going to regulate what I do. That is the height of passing the ball. They could regulate themselves, and they could start by explaining exactly what sources they do use and what sources they could use.

We announced about an hour and a half ago that we're going to offer for all of these companies the kind of audit that we did for Chat version 4, which is we'll take random samples of our misinformation fingerprints, the narratives, and we'll tell these companies internally where they're failing and where they're succeeding – what percent of these false narratives their machines will readily repeat.

KENNEALLY: One of the lessons of the web for the last 20 years is it's given the tools of publishing to everyone – not just to gatekeepers, but to us all. In a way, this is returning to a democratization of media, because you described that the news we read, the information we can, can be curated for us individually. I'd like you to share a little bit more about that vision, that vision of an AI-assisted, AI-boosted newsroom.

CHUA: Look, I don't want to sound like I'm the – I can't remember now, Pollyanna or Cassandra. I don't want to sound like I'm the over-optimist on this. But I do think that there are some good things that will come of this, and I think there are some horrific dangers that we have to be clear about.

I did write a piece a little while back that said, look, one of the problems we have with using chatbots as our search engines is that essentially it summarizes, and it summarizes without citation. What we could be doing – we could simply improve the Google experience dramatically by having – rather than links and little snippets, by summarizing those articles, but also then having the citation or the source. Frankly, if we had something like NewsGuard – and I know, Gordon and Steve, you are doing that – and having a rating on that, you could see how that experience would at the very least allow people to think about what they're getting, as opposed to right now, you ask a question, it gives you a nice, authoritative-sounding answer. That's frankly a slightly worse experience than the current

Google experience, which by itself isn't great, which is a bunch of links that I have to click on to read. So I think you can get to a good point.

I do think, though, speaking to the point of democratization, that again, if you stop thinking about this as an information system, but instead as a language system, there are all sorts of uses that come to mind. One of the things – again, I wrote a piece yesterday about how Microsoft is thinking about integrating this with Excel. Now, I'm happen to love Excel, because I'm a nerd, but it does take some time to learn how to use it. If you had a tool that allowed people to speak essentially in English and have it build spreadsheets for you, again, you can think about the improvements in at least accessibility for people, right – exactly the same way that if you thought about how you used to have to build a website by learning HTML and coding it up yourself, and then a whole bunch of tools came along, and now any of us can throw up a website in five minutes. That's the potential improvement and democratization of access to tools that would otherwise be more complex to use.

It also, of course, adds to the issue of disinformation and more people having access to more publishing platforms or more powerful tools. But to some extent, you can't have it both ways. You can't have a gatekeeping or a gate-kept world where only certain people are allowed to do things, and others don't, because we've had that world. It's not such a great world. And a world where more people can do things – but at the same, that obviously leads to more potential for more damage.

KENNEALLY: Just yesterday, the Future of Life Institute, which is about transformative technology and steering it towards benefiting life and away from risk, issued an open letter. They called for a pause on these giant AI experiments. Elon Musk has signed that letter, Apple co-founder Steve Wozniak, former presidential candidate Andrew Yang. Just in a minute or two, I want to ask the panel if anybody here has signed that letter or thinks it's a good thing to do to sign that letter, or is that letter just part of all the hype that we're experiencing? Anyone want to sign that open letter to pause AI development? They're calling on all AI labs to just stop for six months so we can begin to reflect. And they're also asking policymakers to become more involved. What do you think of that? Perhaps, Tracey Brown –

CHUA: I wish it could happen. I don't think there's any way in hell it will actually happen. But it would be a good idea. Not to be dramatic, the chance of a smart AI machine wiping out humanity is not zero. And it probably isn't going to come to that, but the chance of something really bad happening as it gets more ability to control the interfaces and the devices of the world that we live in I think is – that's a real issue, and we should really think about that.

KENNEALLY: Tracey Brown, what about that letter? Would you sign that, or have you thought about that?

BROWN: I'm hesitant about that, partly because I'm not optimistic that the six months is going to turn a whole lot around. I'd be more interested in hearing the ideas for turning stuff around than the pause. But I am concerned that people's imaginations are firing about, as Gina said, wiping things out. I suspect that the way that we'd wipe stuff out is with banal error rather than with evil intent.

And it does concern me – for example, we live in a situation where major banks aren't completely sure how they're making decisions about who's creditworthy, because it's become a big black box. And we're looking at banks collapsing on a bit of a rumor and a realization that they don't quite know what's going on under the hood. So in a febrile atmosphere like that, where people's pension funds can get wiped out by a sudden collapse of an internal system in a bank, we need to think about what could be the implication of rolling out error. As I say, it could be a mundane thing, but rolling out error, losing a record, destroying a whole bunch of stuff that didn't look important, but suddenly turned out to be – those are the things that concern me about this – just absence of a conversation about any of those things and people in serious positions of authority – government commissioning, for example, in information services and banking and so on who really don't actually understand what the tools are that they're using.

KENNEALLY: Tom Chatfield has written books about all of this – the culture of video games and the nature of political activism. He calls himself a tech philosopher. And in a Twitter thread last week that has 28,000 views, Tom Chatfield noted that AI's bottomless fluency encourages us to treat words as fungible content, more interchangeable vessels for information. We are indeed creating new forms of intelligence utterly unlike our own, Chatfield wrote – mindless, diffuse, abstract intelligence. Something in, but not of our world. Something that deserves curious, skeptical, rigorous, precise engagement.

It's certainly my hope that today's CCC Town Hall, ChatGPT and Information Integrity, has appealed to your curiosity, your skepticism, and your engagement. And if it has, that's entirely because of my guests – Steven Brill and Gordon Crovitz with NewsGuard, Tracey Brown with Sense about Science, Gina Chua with Semafor, and Mary Ellen Bates of Bates Information Services. Thank you all.

Rob Simon of Burst Marketing is our director. Thanks as well to my CCC colleagues, Joanna Murphy Scott, Amanda Ribeiro, Hayley Sund, and Molly Tainter. Stay informed on the latest developments in publishing and research by subscribing to CCC's Velocity of Content blog and podcast.

I'm Christopher Kenneally for CCC. Thanks for joining us. Goodbye for now.

END OF FILE