MISSISSAUGA / ST.GEORGE / SCARBOROUGH

University of Toronto MAGAZINE

How U of T is helping to shape the future of artificial intelligence

Spring 2024

AI EVERYWHERE

Artificial intelligence already seems ubiquitous. In one version of the future, this bodes well: AI will turbocharge progress, lead to new ways to treat disease, warn us of public health threats and even generate new career possibilities. In a darker scenario, it could eliminate entire job categories and fuel a tidal wave of disinformation.

Ensuring the safe and responsible development of AI will be crucial. Read on as we explore how U of T researchers at the Schwartz Reisman Institute for Technology and Society, the Vector Institute and across the three campuses are harnessing the power of AI for good – and helping to shape its future.

ILLUSTRATION BY SEBA CESTARO

HEALING POWER

Why AI could be good news for both patients and our health-care system

About eight years ago, artificial intelligence seemed poised to revolutionize health care. IBM's much-hyped AI system, known as Watson, had rapidly morphed from winning game-show contestant to medical genius, able to provide diagnoses and treatment plans with lightning speed. Around the same time, Geoffrey Hinton, a U of T professor emeritus, famously declared that human radiologists were on their way out.

Now, it's 2024: radiologists are still with us, and Watson Health is not. Have AI and medicine parted company? Quite the opposite, in fact: today, the marriage of disciplines is more vibrant than ever.

At the Temerty Centre for AI Research and Education in Medicine, director Muhammad Mamdani is leading developments that are transforming the field. The centre has more than 1,400 members in 24 universities across Canada and is, Mamdani believes, the largest hub for AI and medicine in the world.

In his role as vice-president of data science and advanced analytics at Unity Health Toronto – a position he held prior to the centre's official launch in 2020 and still retains – Mamdani oversees a team that has created more than 50 new AI solutions, the majority of which are now being deployed.

U of T's combined resources have been essential to the success stories emerging from both institutions. "We have one of the top medical schools in the world," says Mamdani. "As well as highly ranked computer-science, electrical and computer engineering, and statistics departments. So, we've got an incredibly talented pool of researchers."

One of the biggest success stories has been CHARTwatch. The algorithm runs every hour on the hour, analyzing information from patients' electronic records to predict whether the patient's condition will deteriorate and, when the risk exceeds a certain threshold, page the medical team.

CHARTwatch, a Unity Health initiative, has been in operation since 2020 and has been trained on the data of more than 20,000 patients. At the time of its implementation, mortality rates in St. Michael's Hospital (part of Unity Health) were much higher than usual due to COVID-19. But following the deployment of CHARTwatch, the hospital saw a 26 per cent drop in unanticipated mortality compared to pre-pandemic levels. "Peoples' lives are being saved with solutions like this," says Mamdani.

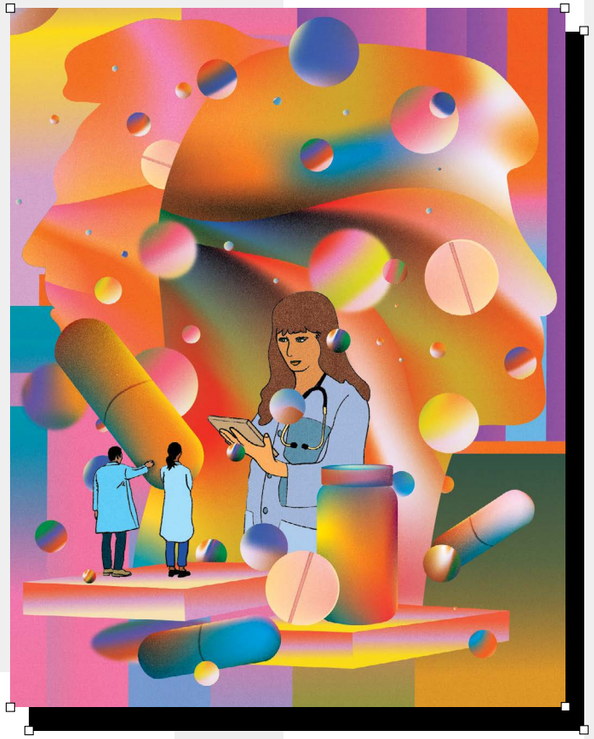
Another algorithm in use, the ED RN Assignment Tool, has reduced the time registered nurses (RNs) spend on scheduling in emergency departments (EDs). "They were struggling with making assignments, because there are all sorts of rules," says Mamdani. The new tool reduces the number of times that any individual nurse is repeatedly assigned to the same role or location within the same series of shifts; it also makes sure that each role is assigned the number of required nurses, each of whom is only assigned to roles or locations that they are qualified to fill.

Since the tool's deployment in 2020, the senior nurse has found that this work can be completed in one minute instead of 90. "With this," says Mamdani, "we're giving time back to clinicians so they can spend it on more valuable activities, such as patient care."

The genesis for these tools is most often clinicians themselves, rather than data scientists, says Mamdani: "We get our ideas from

THE LANGUAGE OF AI WINDING THROUGH THESE PAGES, YOU'LL FIND DEFINITIONS FOR WORDS THAT HAVE TAKEN ON NEW MEANINGS IN THE AGE OF AI.

ALIGNMENT: AN AI SYSTEM IS CONSIDERED IN ALIGNMENT WITH A PERSON OR GROUP WHEN IT APPLIES ITS CAPABILITIES TOWARD ACHIEVING THE INTENDED GOALS OF THAT PERSON OR GROUP. WITH LANGUAGE MODELS, ALIGNMENT IS SOMETIMES USED IN A NARROWER SENSE TO MEAN THE MODELS FOLLOW USER INSTRUCTIONS BUT WON'T OUTPUT HARMFUL CONTENT SUCH AS HATE SPEECH OR ADVICE ON HOW TO COMMIT CRIMES.



people on the ground because they know what the issues are."

The continuing involvement of health-care providers in AI use and development is a key point. About 10 per cent of the Canadian workforce is involved in health care, and, like employees in other fields, many may fear replacement. Mamdani emphasizes that humans must continue to drive AI – not the other way around.

While algorithms have sometimes been shown to outperform clinicians, it isn't always the case. This is why, for the moment, AI is never the sole decision-maker. "For CHARTwatch, for example, we are very firm that it should not decide *for* clinicians, but *with* them. That kind of joint human-AI collaboration is what's driving our reduction in mortality."

To drive this message home, Mamdani says the Temerty Centre for AI Research and Education in Medicine performs an important role as a place where practitioners in the community can learn more about AI, and data scientists can learn about health care. "A huge focus for us is to educate health-care providers – not only about AI's potential to improve the health-care system, but about the challenges." These include ethical considerations around algorithmic gender and racial bias, how well the algorithm performs and how to adopt AI into clinical practice.

"Joint human-AI collaboration is what's driving our reduction in mortality"

"It's very hard to do AI without data," says Mamdani, "and where do you go for it? MIMIC is the world's most publicly used clinical dataset, but researchers need more. They also deal with questions of where to store the data, and whether they have the computing power to do the analyses they want to do."

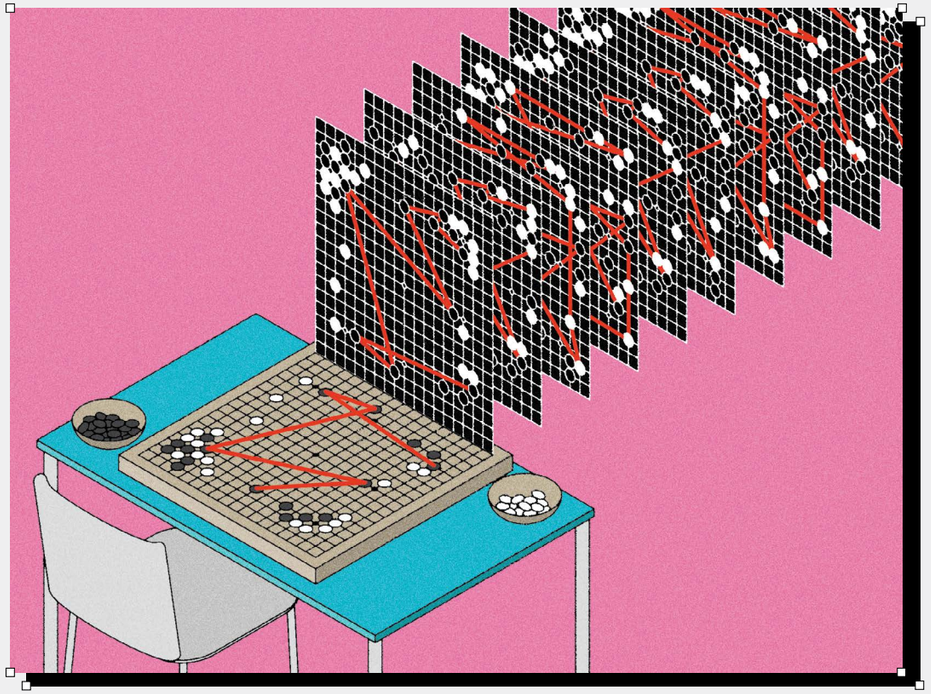
These questions have led the centre to create the Health Data Nexus on Google Cloud, which contains multiple publicly available, large health datasets that community members can access and contribute data to. (Identifiers such as name, address and birthdate are removed.)

What does the future hold for AI in medicine? Mamdani is particularly excited about how it will enable patients to engage in greater self-care: many who might otherwise have been hospitalized will be able to access providers while still at home. "Patients will have monitors and sensors that they can apply themselves, allowing physicians and nurses to videoconference with them and monitor their progress. This could free up hospital beds," he says. "AI may also be able to monitor you. If it detects something wrong, it could flag your provider immediately."

Mamdani is clearly proud of the achievements of both Unity Health Toronto and the Temerty Centre for AI Research and Education in Medicine, which launched less than four years ago. But he is careful to strike a cautious note when discussing the future, so as not to repeat the false promises made years ago.

At the same time, "if we want to live in a society that advances, we have to dream of what's possible," he says. "Educating people about AI's amazing potential, along with its limitations, will lay the foundation for societal acceptance – and the ongoing development of amazing products that we can actually use." **—Cynthia Macdonald**

AI LEARNS EVERYTHING IT KNOWS FROM HUMANS. WILL HUMANS ALSO LEARN FROM AI?



ILLUSTRATIONS BY (LEFT) SEBA CESTARO; (RIGHT) RISTO AVRAMOVSKI

In 2016, an AI program called AlphaGo made headlines by defeating one of the world's top Go players, Lee Sedol, winning four games of a five-game match. AlphaGo learned the strategy board game by studying the techniques of human players, and by playing against versions of itself. While AI systems have long been learning from humans, scientists are now asking if the learning could go both ways. Can we learn from AI?

Karina Vold, an assistant professor at U of T's Institute for the History and Philosophy of Science and Technology, believes we can. She is studying how humans can learn from technologies such as the neural networks that underlie today's AI systems.

"In the case of Go, professional players learn through proverbs such as 'line two is the route to defeat,' or 'always play high, not low,'" says Vold, who works at the intersection of philosophy and cognitive science, and is affiliated with the Schwartz Reisman Institute for Technology and Society. Those proverbs can be useful, but they can also be limiting, impeding a player's flexibility. AlphaGo, meanwhile, gleans insights – a term that Vold believes is appropriate – from digesting enormous volumes of data. "Because AlphaGo learns so differently, it did moves that were considered very unlikely for a good human player to make," Vold says.

A key moment occurred in the second game, on the 37th move, when AlphaGo played a move that took everyone – including Sedol – by complete surprise. As the game went on, however, move 37 proved to be a masterstroke. Human Go players "are now studying some of the moves that AlphaGo made and trying to

come up with new sorts of proverbs and new ways of approaching the game," says Vold.

Vold believes the possibility of humans learning from AI extends beyond game playing. She points to AlphaFold, an AI system unveiled by DeepMind (the same company behind AlphaGo) in 2018, that predicts the effects of proteins based on their structure. Proteins are made up of sequences of amino acids, which can fold and form complex 3D structures. The protein's shape determines its properties, which in turn determine its potential efficacy in new drugs to treat diseases. Because proteins can fold in millions of different ways, however, it is impossible for human researchers to work through all the combinations. "This was a long-standing grand challenge in biology that had been unsolved," says Vold, but in which AlphaFold "was able to make great advances."

Even in cases where humans may have to rely on the sheer computing power of an AI system to tackle certain problems – as with protein folding – Vold believes artificial intelligence can guide human thinking by reducing the number of paths or conjectures that are worth pursuing. While humans may not be able to duplicate the insights an AI model makes, it is possible "that we can use these AI-driven insights as scaffolding for our own cognitive pursuits and discoveries."

In some cases, Vold says, we may have to rely on the "AI scaffolding" permanently, because of the limitations of the human brain. For example, a doctor can't learn to scan medical images the same way that an AI processes the data from such an image; the brain and the AI are just too different. But in other cases, an AI's outputs "might serve as cognitive strategies that humans can internalize [and, in so doing, remove the 'scaffolding']," she says. "This is what I am hoping to uncover."

Vold's research also highlights the issue of AI "explainability." Ever since AI systems started making headlines, concerns have been raised over their seemingly opaque workings. These systems, and the neural networks they employ, have often been described as "black boxes." We may be impressed by how quickly they appear to solve certain kinds of problems, but it might be impossible to know how they arrived at a particular solution.

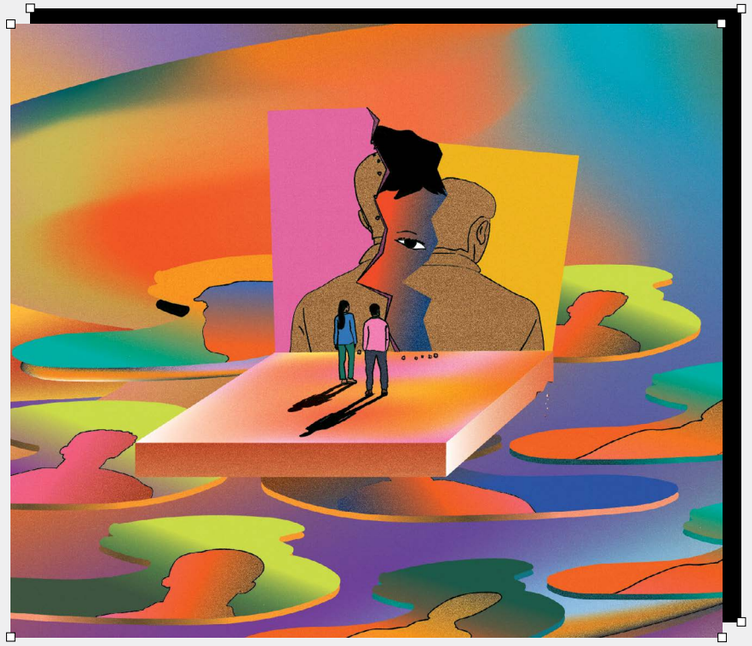
Vold believes it may not always be necessary to understand exactly how an AI system does what it does in order to learn from it. She notes that the Go players who are now training on the moves that AlphaGo made don't have any inside information from the system's programmers as to why the AI made the moves it did. "Still, they are learning from the outputs and incorporating the moves into their own strategic considerations and training. So, I believe that at least in some cases, AI systems can function like black boxes, and this will be no hindrance to our learning from them."

Yet there may still be situations where we won't be satisfied unless we can see inside the black box, so to speak. "In other cases, we may need to understand how the system works to really learn from it," she says. Trying to distinguish cases where explainability is crucial from those where a black box model is sufficient "is something I'm still thinking about in my research," says Vold. **—Dan Falk**

↓ Professor Karina Vold

PHOTOGRAPH BY DUANE COLE; ILLUSTRATION SEBA CESTARO

⇨ ⇨ ⇨ **PROMPT:** THE INPUT GIVEN, OFTEN BY A HUMAN USER, TO AN AI MODEL TO GENERATE A RESPONSE. THE PROMPT FOR A LANGUAGE MODEL SUCH AS CHATGPT, OR A TEXT-TO-IMAGE MODEL SUCH AS DALL-E, COULD BE A SENTENCE, QUESTION, COMMAND OR OTHER TEXT. THE PROMPT GENERALLY DEFINES A TASK AND IMPLICITLY PROVIDES CONTEXT FOR WHAT PARTS OF THE MODEL'S TRAINING DATA ARE RELEVANT TO FULFILLING IT. ⇨ ⇨ ⇨



THE AGE OF DECEPTION

AI is generating a disinformation arms race. The window to stop it may be closing

In an apparent interview with the talk show host Joe Rogan a year ago, Prime Minister Justin Trudeau denies he has ever appeared in blackface, responds to rumours that Fidel Castro was his father, and says he wishes he had dropped a nuclear bomb on protesters in Ottawa.

The interview wasn't real, of course, and was apparently intended to be humorous. But the AI-generated voice of Trudeau sounded convincing. If the content had been less absurd it would have been difficult to distinguish from the real thing.

The video highlights the growing danger that artificial intelligence could usher in a new era in disinformation – one that will make it easier than ever for malign actors to spread propaganda and fake news that seems authentic and credible. Recent advances in generative AI have made it much easier to create all sorts of convincing fake content – from written stories to replicated voices and even fake videos. And as the technology gets cheaper and more widespread, the danger increases.

"It's probably one of the things that I'm most worried about right now," says Ronald Deibert, director of the Citizen Lab at the Munk School of Global Affairs and Public Policy. "I think it's going to create all sorts of chaos and havoc, and amplify a lot of the problems that we're seeing around misinformation and social media," he says.

AI tools such as ChatGPT allow people to write articles about specific subjects in a particular style. For instance, U.S. researchers were able to get the program to draft convincing essays arguing that the school shooting in Parkland, Florida, was faked, and that COVID-19 could cause heart problems in children. "You can simply plug in a prompt and the entire article can be created. That makes it just so much easier," Deibert says. "It's hard to discern if something is fake or authentic."

Impersonating a voice is also straightforward. The people who made the Trudeau fake interview said they used a service called ElevenLabs. The company's website offers the ability to generate a realistic human voice from a typed script, and also has the option of "cloning" a voice from a recording.

A technology like this may have been used in January during the New Hampshire presidential primaries, when a robocall in the voice of President Joe Biden urged Democrats not to vote. The New Hampshire Attorney General's office said the recording seemed to use an artificially generated voice.

Perhaps even more concerning are deepfake videos, which can be made to show a lookalike version of a real person doing or saying almost anything. For instance, a video that appeared last year seemed to depict Hillary Clinton on MSNBC endorsing then-Republican presidential candidate Ron DeSantis. Although the

TOKEN: A SHORT SEQUENCE OF CHARACTERS, SUCH AS LETTERS, PUNCTUATION MARKS AND SPACES THAT SERVES AS THE FUNDAMENTAL BUILDING BLOCK FOR ANALYZING TEXTUAL DATA WITHIN AI SYSTEMS. FOR CHATGPT, 1,000 TOKENS IS ABOUT 750 WORDS. SOMETIMES USERS ARE BILLED BASED ON HOW MANY TOKENS APPEAR IN THE INPUTS THEY GIVE TO, AND OUTPUTS THEY RECEIVE FROM, AI LANGUAGE MODELS.

face seemed slightly rubbery, the video was fairly convincing – until the end, when Clinton says, "Hail, Hydra!" – a reference to an evil organization from Marvel comics and movies.

The stakes can be high. In 2022, a deepfake video of Ukrainian President Volodymyr Zelenskyy appeared to show him urging Ukrainian soldiers to put down their arms and surrender.

In the recent past, creating forged documents, photos or articles took considerable time and effort. Now, generating synthetic media is simple, widely available and cheap. One researcher, whose work is widely known but who has not disclosed his identity, built and demonstrated an AI-enabled platform called Countercloud that was able to create a disinformation campaign – complete with fake news articles and extensive social media support – with only a few prompts. "So what you have is the means to generate authentic, credible-looking content with the push of a button," Deibert says. This substantially lowers the barriers for malicious actors who want to wreak havoc.

Deibert and his colleagues at the Citizen Lab have documented several sophisticated disinformation campaigns on social media. They recently released a report by researcher Alberto Fittarelli on an effort they call Paperwall, in which at least 123 websites run from within China impersonate legitimate news sites from around the world, running stories favourable to Beijing. Previous work by the lab has uncovered sophisticated disinformation campaigns run on behalf of Russia and Iran.

Deibert isn't the only one raising the alarm about AI and disinformation. Publications from the *New York Times* to *Foreign Affairs* have run articles about the problem – and possible solutions. Among them are technical approaches, such as "watermarks" that allow users to see if information has been generated by an AI, or artificial intelligence programs that are capable of detecting when another AI has created a deepfake. "We will need a repertoire of tools," Deibert says – "often, the same tools the bad actors are using."

Social media companies also need to devote more resources to detecting and eliminating disinformation on their platforms. This might require government regulation, he says, though he acknowledges that this comes with the risk of government overreach. He also calls for increased regulation around ethical use of and research into AI, noting this would apply to academic researchers as well.

But Deibert thinks a broader solution is needed, too. A big part of the problem, he says, are social media platforms that depend on creating extreme emotions in users to keep them engaged. This creates a perfect breeding ground for disinformation. Convincing social media companies to turn the emotional volume down – and educating citizens to be less prone to manipulation – may be the best long-term solution. "We need to rethink the entire digital ecosystem to deal with this problem," he says. **—Kurt Kleiner**

"We need to rethink the entire digital ecosystem to deal with this problem"

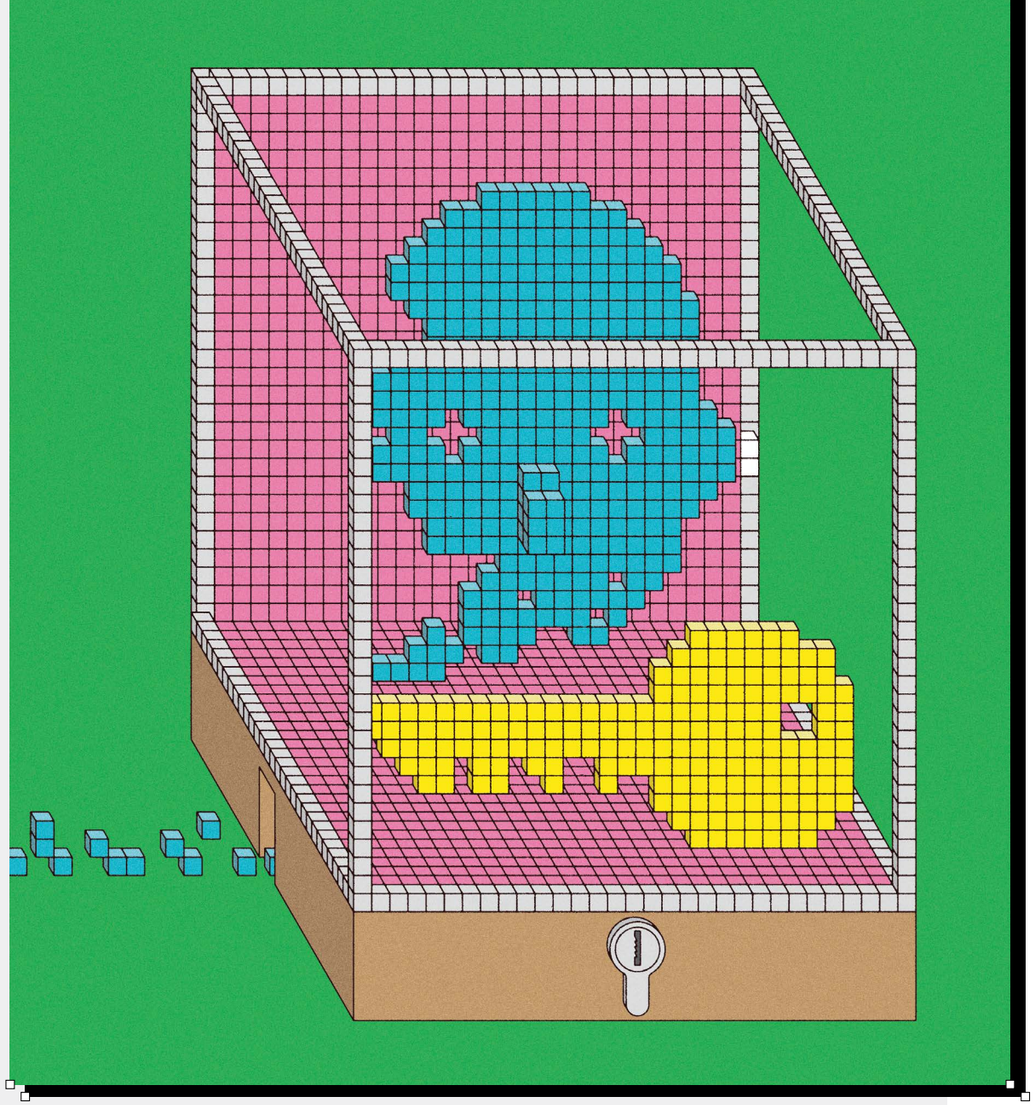


ILLUSTRATION BY RISTO AVRAMOVSKI

SAFETY FIRST

AI has developed faster than anyone thought. Will it serve humanity's best interests?

It's no secret that artificial intelligence has unleashed a variety of potential dangers. AI systems, for example, can be used to spread misinformation; they can perpetuate biases that are inherent in the data they have been trained on; and autonomous AI-empowered weapons may become commonplace on 21st-century battlefields.

These risks are, to a large extent, ones that we can see coming. But Roger Grosse, an associate professor of computer science at U of T, is also concerned about new kinds of risks that we might not perceive until they arrive. These risks increase, Grosse says, as we get closer to achieving what computer scientists call artificial general intelligence (AGI) – systems that can perform a multitude of tasks, including ones they were never explicitly trained to do. "What's new about AGI systems is that we have to worry about the risk of misuse in areas they weren't specifically designed for," says Grosse, who is a founding member of the Vector Institute for Artificial Intelligence and an affiliate of U of T's Schwartz Reisman Institute for Technology and Society.

Grosse points to large language models, powered by deep-learning networks, as an example. These models, which include the popular ChatGPT, aren't programmed to produce any particular output; rather, they analyze massive volumes of text (and images and videos), and respond to prompts by stringing together single words based on the likelihood of that word occurring next in the data it was trained on. While this may seem like a haphazard way of building sentences, systems such as ChatGPT have nonetheless impressed users by writing essays and poems, analyzing images, writing computer code and more.

And they can catch us by surprise: Last year, Microsoft's Bing chatbot, powered by ChatGPT, told journalist Jacob Roach that it wanted to be human, and was afraid of being shut down. For Grosse, the challenge is trying to determine what sparked that output. To be clear, he doesn't think the chatbot was actually conscious, or actually expressing fear. Rather, it may have come across something in its training data that led it to say what it said. But what was that something?

TO TACKLE THIS problem, Grosse has been working on techniques involving "influence functions," which are designed to deduce what aspects of an AI system's training data led to a particular output. For example, if the training data included popular sci-fi stories, where tales of conscious machines are ubiquitous, then this could easily lead an AI to make statements similar to those found in such stories.

He notes that an AI system's output may not necessarily be copied word-for-word from the training data, but rather may be some variation on what it's encountered. They can be "thematically similar," Grosse says, which suggests that the AI is "emulating" what it has read or seen and performing "a higher level of abstraction." But if the AI model develops an underlying motivation, this is different. "If there were some aspect of the training procedure that is rewarding the system for self-preservation behaviour, and this is leading to a survival instinct, that would be much more concerning," says Grosse.

Even if today's AI systems aren't conscious – there's "nobody home," so to speak – Grosse believes there could be situations where it makes sense to describe an AI model as having "goals." Artificial intelligence can surprise us by "behaving as if it had a goal, even though it wasn't programmed in," he says.

These secondary or "emergent" goals crop up in both human and machine behaviour, says Sheila McIlraith, a professor in the department of computer science and associate director and research lead at the Schwartz Reisman Institute. For example, a person who has the goal of going to their office will develop the goal of opening their office door, even though it wasn't explicitly on their to-do list.

The same goes for AI. McIlraith cites an example used by computer scientist Stuart Russell: If you tell an AI-enabled robot to fetch a cup of coffee, it may develop new goals along the way. "There are a bunch of things it needs to do in order to get that cup of coffee for me," she explains. "And if I don't tell it anything else, then it's going to try to optimize, to the best of its ability, in order to achieve that goal. And in doing that, it will set out other goals, including getting to the front of the line of the coffee store as quickly as possible, potentially hurting other people because it was not told otherwise."

Once the AI model is developing and acting on goals beyond its original instructions, the so-called "alignment" issue becomes paramount. "We'd like to make sure the goals of AI are in the interest of humanity," says Grosse. He adds that it makes sense to say that an AI model can reason, if it works through a problem step by step, the way a human would. The fact that AI can seemingly solve difficult problems is something that not long ago would have seemed miraculous. "That's why we're in a different situation from where we were a few years ago," says Grosse, "because if you asked me in 2019, I would have said deep learning can do a lot of amazing things, but it can't reason."

For Grosse, the acceleration of AI technology, as illustrated by the capabilities of today's large language models, is cause for concern – and a reason to refocus his research on safety. "I'd been following the discussions of AI risk for a long time," he says. "I'd sort of bought into the arguments that if we had very powerful AI systems, it would probably end badly for us. But I thought that was far away. In the last few years, things have been moving much faster."

Even if the frightening scenarios depicted in the Terminator movie franchise are more Hollywood than reality, Grosse believes it makes sense to prepare for a world in which AI systems come ever-closer to possessing human-level intelligence, and have some measure of autonomy. "We need to worry about the problems that are coming," says Grosse, "where more powerful systems could actually pose catastrophic risk." **—Dan Falk**

⇨ ⇨ **GUARDRAIL:** A GUIDELINE, CONSTRAINT OR SAFETY MEASURE PLACED ON AN AI MODEL TO MAKE SURE IT OPERATES WITHIN PREDEFINED BOUNDARIES AND DOESN'T DO BAD THINGS, SUCH AS OUTPUT HARMFUL CONTENT OR EXHIBIT BIAS OR DISCRIMINATION. ⇦ ⇦

ILLUSTRATION BY SEBA CESTARO

TUNING INTO TOMORROW

AI can help musicians compose and create new sounds. Is it just another music-making tool – or something else?

About a year ago, Stephen Brade started noodling around with a guitar composition. He set the unfinished piece aside, but returned to it when he realized that an AI-powered synthesizer he was developing might help him find the swelling yet spacious sound he was striving for.

Brade, a master's student in the computer science department at U of T, is the creator of SynthScribe, a research project that aims to make synthesizers more user-friendly by allowing musicians to shape sounds through text and audio inputs rather than complex manual adjustments.

To demonstrate SynthScribe, Brade first walks a visitor through the constellation of buttons and knobs that musicians must learn

how to precisely manipulate to design synthesizer sounds. One shortcut in this time-consuming process is to draw from vast libraries of premade sounds, but it can be difficult to come up with the exact search terms to match the sound in your head.

There's a disconnect, Brade says, between the technical jargon often tagged to sounds ("saw wave," "attack time," "low-pass filter") and the subjective everyday language we use to describe them ("warm," "gritty," "dreamy").

SynthScribe uses advanced AI to bridge this gap, making it easy for users to find, change and create synthesizer sounds using both descriptive words and audio clips.

Brade types "sound of the void" into the system's search bar and a hum of white noise whooshes through the speakers. He asks the system to make a flute sound "harsher" and it adjusts to a more bracing pitch. There's also a feature that can blend sounds together to create brand new ones. For his own piece, Brade landed on a raspy, resonant sound that he feels adds to the song's lilting melancholy.

Brade says his own creative experience was consistent with what musicians who were asked to try SynthScribe told the researchers. Many of the artists surveyed highlighted the system's ability to help them think outside the musical box, generating sounds they did not expect but were still pleased to hear. "There's a lot of potential for really, really new music," he says.

Brade is among those who are excited about AI's potential to unleash a new wave of musical experimentation by creating never-before-heard sounds, streamlining production methods and reducing the technical barriers to creative expression.

But the rise of generative AI has also sowed discord in the music community. There are fears of unscrupulous actors using AI to "clone" singers' voices and of job losses in sound production. (Brade himself is wary of the potential for musicians to be exploited.) Some even fear that AI will strip music of its soul.

As a coder who enjoys composing, Brade says the outlook might depend on whether AI programs are designed to serve musicians as artistic collaborators or supplant them as all-in-one, automated music creators. So far, Brade thinks humans still have the edge. "Generative models tend to create music that sounds derivative," he says. "This is less likely to be the case with a composer or musician who is trying to push boundaries."

Gregory Lee Newsome, an assistant professor, teaching stream, in U of T's music technology and digital media graduate program, sees AI as simply the latest – but most powerful – example of technology's influence on the trajectory of music.

Newsome, who provided technical support on SynthScribe and co-authored the preprint paper, says artists have always made use of new tools. But he worries that generative AI might be qualitatively different than any innovation that's come before: "It's so powerful that it may not require human intervention at all."

Technologies such as Stability AI's Stable Audio and Google DeepMind's Lyria already allow users to compose music in a variety of genres and styles without having to play a note.

Meanwhile, an AI-generated "collaboration" between Drake and The Weeknd that went viral last year has raised alarm about vocal clones, spurring mixed reactions within the music industry.

Last fall, Universal Music Group and other music publishers sued AI company Anthropic over allegations that its chatbot Claude copies and distributes copyrighted song lyrics. It's one of several similar lawsuits filed by copyright owners – including writers, visual artists and the *New York Times* – claiming their content was improperly used to train AI models.

At the same time, however, a number of music industry players are looking to get in on the AI action. Universal, for example, teamed up with YouTube to guide its approach to AI-generated music, while hitmakers including Charlie Puth and T-Pain have lent their voices to a Lyria-powered experiment on YouTube Shorts.

People often experience music emotionally, Newsome says, and many fans form deep personal attachments to their favourite artists. This essential element of human connection could prove difficult for AI to replicate, he suggests.

"If people like the tune, the voice, the lyrics, the beat, will they care about who – or what – made it?"

But Newsome says he still fears for his students, most of whom remain committed to traditional modes of music production, as AI threatens to siphon off the already limited revenue streams available to musicians.

"It's a little mysterious what's going to happen, but I would not be surprised if this is a sea change for a lot of music production," he says. "The test may be: if people like the tune, the voice, the lyrics, the beat, will they care about who – or what – made it?" **—Adina Bresge**

ILLUSTRATION BY SEBA CESTARO

JAILBREAK: COMPANIES PUT IN A LOT OF EFFORT TO GET THEIR LANGUAGE MODELS NOT TO SAY THE F-WORD, PRODUCE HATE SPEECH, OR TELL YOU HOW TO BUILD DANGEROUS WEAPONS. BUT THERE ARE WAYS TO DESIGN PROMPTS TO "TRICK" THE MODEL INTO OUTPUTTING PRECISELY THE CONTENT THAT ITS DESIGNERS HAD TRIED (BY USING GUARDRAILS) TO PREVENT. THESE ARE CALLED JAILBREAKS.

SYCOPHANCY: LANGUAGE MODELS SUCH AS CHATGPT ARE TRAINED TO GET APPROVAL FROM THEIR CURRENT USER. THIS CAN LEAD TO SYCOPHANCY, WHERE THE AI RESPONDS WITH ANSWERS THAT SEEM TO MATCH THE USER'S BELIEFS, REGARDLESS OF THE TRUTH. THE AI MIGHT RESPOND DIFFERENTLY, FOR EXAMPLE, DEPENDING ON WHETHER IT THINKS YOU'RE A LIBERAL OR A CONSERVATIVE. –AI GLOSSARY BY TORYN KLASSEN (POSTDOCTORAL FELLOW) AND LEV MCKINNEY (MASTER'S STUDENT), OF U OF T COMPUTER SCIENCE.

A SENTINEL FOR GLOBAL HEALTH

AI is promising a better – and faster – way to monitor the world for emerging medical threats

In late 2019, a company called BlueDot warned its customers about an outbreak of a new kind of pneumonia in Wuhan, China. It wasn't until a week later that the World Health Organization issued a public warning about the disease that would later become known as COVID-19.

The scoop not only gained BlueDot a lot of attention, including an interview on *60 Minutes*, it also highlighted how artificial intelligence could help track and predict disease outbreaks.

"Surveillance and detection of infectious disease threats on a global scale is a very complex endeavor," says Kamran Khan, a professor at U of T's department of medicine, a clinician-scientist at St. Michael's Hospital and the founder of BlueDot. His solution is to use AI to help sort through vast amounts of information, and tag data of potential interest for the company's human experts to evaluate.

"The metaphor of the needle in the haystack is the right one. We're building a very big, increasingly comprehensive haystack. But identifying what is anomalous or unusual is the key, because many outbreaks appear around the world every day, yet the vast majority of them are limited in scale and impact."

BlueDot founder Kamran Khan →

Khan's career as a doctor is largely defined by big disease outbreaks. He was doing a fellowship in New York in 1999 when West Nile virus hit, and was there in 2001 when anthrax spores were mailed to members of Congress and the media, killing five. He moved to Toronto just months before the SARS outbreak of 2003. "Having seen three infectious disease emergencies in four years was an indication to me that in my career we probably were going to see more of these," he says.

The heart of BlueDot's method was outlined in a paper published six months before COVID hit. The company uses a database of news stories, created by Google and collected from 25,000 sources in 100 languages around the world – far too much for humans to sort through. Instead, an AI model trained by the company sorts through the stories and flags those that seem most likely to be about a disease outbreak of interest. To develop the system, the team ran their program retrospectively on the 12-month period from July 2017 to June 2018, and compared their results with official World Health Organization (WHO) reports for the same period.

In the paper, the researchers explain that online media covered all but four of the 37 disease outbreaks identified by the WHO, and that their system flagged all but two of those reports. Although the system missed some of the outbreaks, the ones it did detect it found much earlier than the WHO – an average of 43 days before an official announcement.

Since that 2019 paper, BlueDot has supplemented news sources with information from government health websites, reports from the medical and health communities and reports from their own clients. "We use the internet to detect early signals that something unusual is going on in a community, even before it is being officially reported," Khan says, noting that government information is sometimes delayed – or suppressed for political reasons.

In their paper, the researchers used historical data to show that their system could work in theory. By the end of the year, when BlueDot detected COVID-19, they were able to show they could also beat official reports in real time.

More recently, BlueDot alerted clients to an outbreak of Marburg virus, which is similar to Ebola, in Equatorial Guinea in February 2023. WHO officials confirmed the outbreak with blood testing, and sent medical experts and protective equipment to help contain the disease. "We can innovate in a way that helps governments and other organizations move more quickly, when time is of the essence," Khan says.

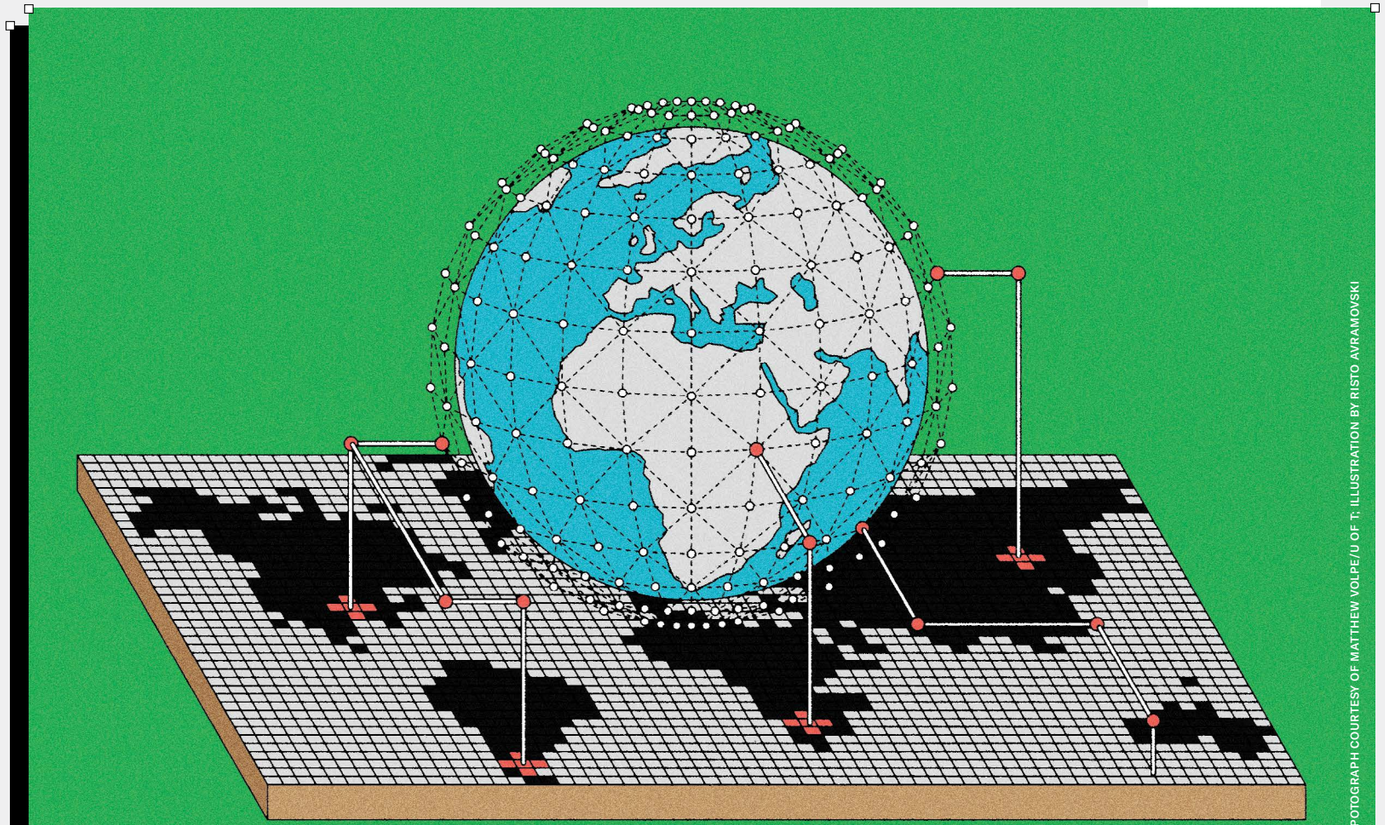
BlueDot isn't the only organization using AI to help keep track of outbreaks, Khan says. The WHO has a system, and there are other start-ups and non-profits experimenting with the method. Khan says BlueDot offers extra value by making the information easy for governments and private clients to access, use and act upon – and by providing further analysis. "Our belief is that epidemics are a whole-of-society problem, which means that organizations across sectors need to be empowered to do their part," he says. Since starting more than a decade ago, BlueDot has attracted 30 clients in 24 countries, both public and private. Its government clients represent 400 million people.

Khan notes that improvements in AI since 2019 are opening new possibilities for using the technology to analyze and report information. Sorting through the data generated by the system can be a time-consuming task for a human. Khan says AI is getting better at generating text and visuals such as infographics, and humans are no longer needed to create routine reports with summaries, charts and simple analysis. BlueDot is also using a new interface that allows people to ask questions about disease outbreaks in everyday language. In the past, working with the system required some computer coding skills, he says.

Soon, the company would like to be able to take its audience into consideration – for instance, writing one kind of report for a doctor, and another for a policymaker. "Generative AI is opening the door for us to be able to communicate insights to a diverse set of audiences at a large scale," Khan says.

Rather than replacing humans, he thinks AI will complement them, allowing teams of experts to analyze and make decisions about much more data than they could otherwise. **—Kurt Kleiner**

PLUS: COULD AI HELP REDUCE INCOME INEQUALITY? READ MORE AT MAGAZINE.UTORONTO.CA.



POTOGRAPH COURTESY OF MATTHEW VOLPE/U OF T; ILLUSTRATION BY RISTO AVRAMOVSKI