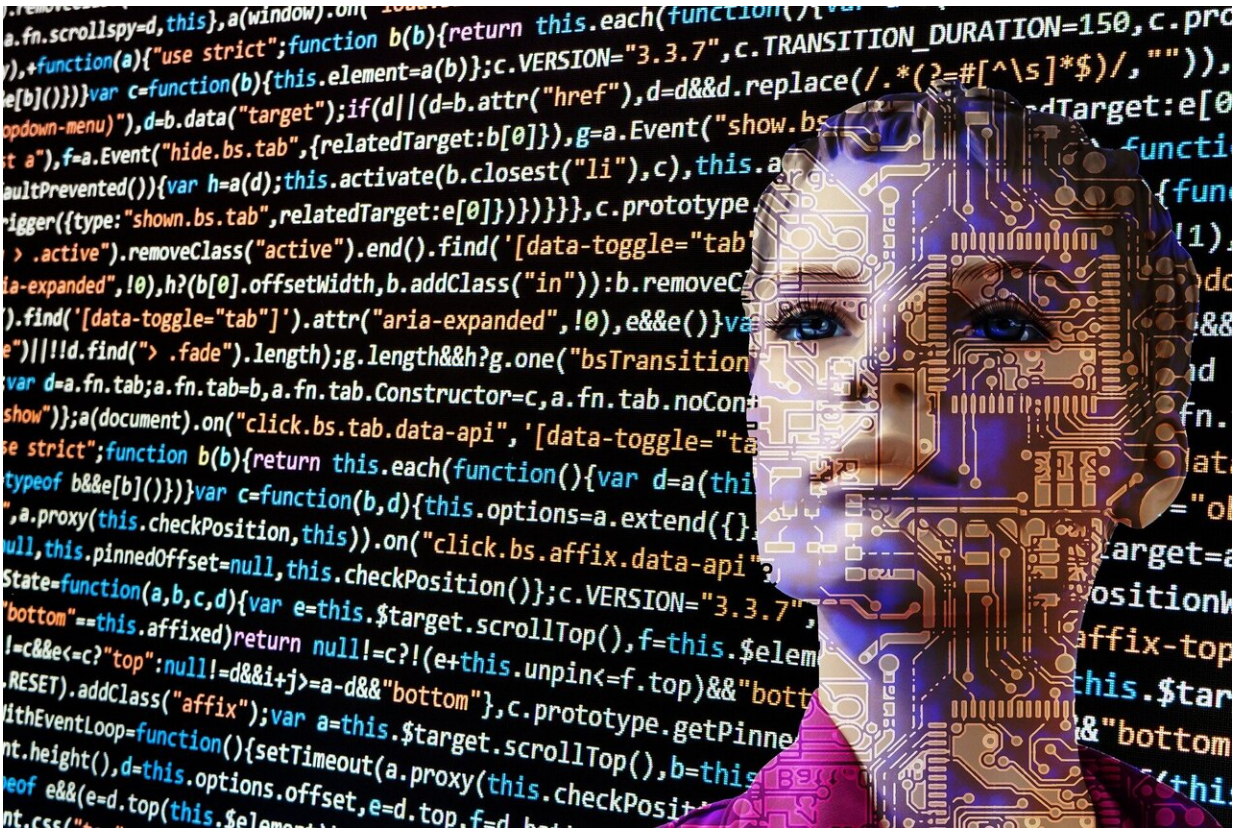
How should AI depict marginalized communities? Technologists look to a more inclusive future

June 26 2024, by Alexander Johnson



Credit: CC0 Public Domain

As artificial intelligence becomes more sophisticated and more capable of closely depicting reality, researchers at Carnegie Mellon University's

Human-Computer Interaction Institute (HCII) are working to ensure that the outputs of large language models are representative of the communities they reference.

This work is the primary focus of William Agnew. He is a Carnegie Bosch postdoctoral fellow and one of the leading organizers of Queer in AI. Alongside Carnegie Mellon, Queer in AI is a member of the National Institute of Standards and Technology's AI Safety Institute Consortium, which seeks to advance the trustworthiness and safety of AI systems.

"Researchers, corporations and governments have long and painful histories of excluding marginalized groups from technology development , deployment and oversight," Agnew and the other organizers of Queer in AI wrote in their paper on AI risk management. "As a result, these technologies are less useful and even harmful to minoritized groups."

Since starting to work with this organization approximately eight years ago, Agnew has used his expertise to analyze the integrity of training datasets for large language models. Through his work, he helps AI developers identify and overcome biases across mediums—generated text, images, voice and music—with the end goal of helping technology be more equitable in its application.

"These audits really have the goal of asking, are they representative?" Agnew said. "Are they inclusive or are they biased? Do they contain toxic stereotypes? Are they taking people's intellectual property or other work without their permission? Do communities not want to be in these datasets?"

Answering all of these questions requires diving into the data behind the content. "By understanding the dataset, we can really understand what's going to happen in all the downstream models," he said.

The ultimate goal of this work is to empower those who have previously been left out of discussions around privacy and security in the implementation of AI.

"A lot of communities want to have control over their data and their representations. They don't want companies to decide how they are represented in media or AI. They want to control that," Agnew said. "It's valid and important. Marginalized groups have had decades if not centuries of stereotypes, caricatures and misrepresentation in media."

Agnew also explained that anyone who creates content that is ultimately seen online—not just marginalized communities—would benefit from the ability to opt out of inclusion in these datasets, giving the example of the increasing trend in journalism to train models on the writing of authors without their consent.

How does representation in the AI age differ from the past?

Until recently, questions of representation and belonging—most notably, how marginalized communities depict and perceive themselves—have fallen into the domain of traditional artists, community leaders and historians.

Harrison Apple, founder of the Pittsburgh Queer History Project and associate director of the Frank-Ratchye STUDIO for Creative Inquiry in the College of Fine Arts, noted that the historical archive itself plays a role in recording events and situations as "technologies of belonging"—tools which identify members of a group through shared experience. For archivists, this experience takes the form of a shared and localized past. For technologists and social media users, it is an immediate and globally-accessible present.

Individuals in both contexts are often used as a point of reference to identify or demarcate their entire community, even if they cannot consent to it. "Setting out to form a community is such a fraught project. Community is deployed by whoever wields the word—to protect, to destroy—but it is always a circumscribing concept," Apple said.

Apple identified the ability to use imagery to shape discourse around a topic as a privilege in their article critiquing community archives, titled "I Can't Wait for You to Die." For instance, archival exhibitions often focus on those who are no longer alive with the goal of helping modern-day onlookers develop a sense of community and identity. However, because late individuals cannot approve of the use of their stories or likenesses, questions of ethics can arise for those given personal material.

Apple's solution to this problem as a public historian is the MS '89 screening series, which showcases donated archival LGBTQ+ tapes with their creators in attendance. This approach to archival work is intended to bridge the gap between past and present, encouraging members of local communities to take an active role in defining themselves.

They added that archivists do have a responsibility to invest in innovative, potentially empowering ideas, but that any technology centered on representation should be treated with caution.

"I do not believe that any technology is inherently liberatory," Apple said. "It only has the potential to become part of making our distinct mission more elegant and far-reaching. Before you can change the world, you have to figure out what it is you're asking for."

MS '89, Apple explained, presents a community-based solution to a community-based problem. "In my case, I want to get people in a room together to understand that what changed my life was not getting the

tapes, but watching them with the donors," Apple added. "This is a big part of how I think about programming exclusively archival video footage from nightclubs from before the time of Web 2.0. The tapes weren't made for a peer-to-peer digital public, and we can't reach back in time to make that decision for them."

The combined force of generative AI and social media presents a similar dilemma to the one community archivists are met with: Nearly everything generated and shared online is made exclusively for—and out of—a peer-to-peer public that cannot always consent to it.

The questions of identity and consent become more immediate as individuals can request a depiction of a marginalized person on demand, often at the click of a button.

How does technology typically shape representation?

The story of AI and representation today is inseparable from the role played by social media in recent history.

"Queer communities have a very complicated relationship with online spaces, and especially public online spaces," Agnew said. "On one hand, they are vital community bases for queer people, especially since many of us start out without having any particular queer community at a young age. It's just us and what we could find on Google or Reddit, and these were very valuable, often life-changing or life-saving connections and relationships."

Jordan Taylor, a Ph.D. student at Carnegie Mellon's HCII, studies how marginalized people leverage technology as well as how technology designers and researchers think about marginalized people. He is advised by Assistant Professor Sarah Fox, who leads the Tech Solidarity Lab, and Associate Professor Haiyi Zhu, who leads the Social AI Group.

Taylor's recent research includes an examination of online communities on social media platforms like Reddit, and how they respond to an issue known as "hermeneutical injustice"—the historical inability of a marginalized group to be able to understand themselves due to external societal restrictions.

When looking at these spaces, he found that the digital environment created a unique opportunity for users to interact and see themselves reflected.

"I was looking at the subreddit r/bisexual and trying to understand what people are doing in this community," Taylor said. "We found that people are constructing a particular way of understanding themselves in the world. This includes things like developing ingroup language and ingroup stereotypes. It's constructing these ways to classify themselves and understand how bisexuality is situated in the broader world."

However, this ability for communities to gather and build identity in a digital context is often complicated—and, in many cases, hindered—by the pre-existing motivations and frameworks of technology companies. "That kind of ingroup difference is often flattened and erased when we talk about the design of technology," Taylor said.

In his recent research, Taylor has pivoted to examining the ever-changing relationship between AI-generated content and the marginalized communities interfacing with it online, with a particular focus on LGBTQ+ artists' engagement with generative models like DALL-E 3, Midjourney and Stable Diffusion.

For instance, a site's protocols may identify content by or for LGBTQ+ individuals as harmful or inappropriate. This is often the case for image generators, Taylor said, which flag certain outputs as containing explicit imagery, sometimes inconsistently and with little regard for the context

around the input.

Wholesale erasure of content without considering its cultural significance has been a long-running problem with content moderation algorithms that remains despite recent innovations, he said.

"Marginalized communities often use—and have a long history of using—technologies that were not necessarily designed with them in mind, or where there kind of isn't a particular user in mind at all. You end up modulating to the norm and oftentimes that is a white, straight, wealthy Western norm," Taylor said. "That is the gaze through which we're understanding these groups."

Agnew, whose postdoctoral research is currently focused on music datasets, said that gaze tends to be utilized to aggregate data. "The preliminary results for our study indicate that the biases we see everywhere in AI are also present in these datasets: Men are mentioned much more often than women. White people are mentioned more often than other racial and ethnic groups. Mentions of queer people and other marginalized communities tend to be more negative. This leads to disparate performance in the downstream models."

How is generative AI shaping the future of representation?

As identities are amalgamated in training datasets, questions around how individuals can identify with content or consent to certain representations become more complex, especially if these representations are tailored to external tastes and standards. Apple, in their article, cites authors who discuss how depictions of marginalized groups in Hollywood and Congress have left much to be desired by those they are meant to represent. They say that Silicon Valley (and the tech

world in general) could now reasonably be added to this list.

Generated content relies on previous products of the human imagination, and inherits the biases of the source content's creators. For Apple, these biases and the increasing availability of generative content has implications for archival work as well.

"I work with an archive that may very well have been scraped in pieces into a dataset of images if they were ever captured for use," Apple said. "I have yet to see evidence that it is able to imagine or create something people don't already want to see."

This biased idealization is an issue that Taylor has noticed during his own ongoing research.

"A lot of the artists in my study mentioned that all of the art that they were generating using DALL-E 3 was very finished, polished and symmetrical. It kind of has these corporate aesthetic values embedded in it, but people were struggling to find ways to integrate these very polished images into their artistic practice," he said.

Is generative AI harmful to marginalized communities?

Carnegie Mellon's commitment to safety, responsibility and equity is a guiding principle, approach and component across all areas of AI research at the university.

At CMU, scientists and engineers work with philosophers, artists, economists, ethicists, social scientists and policy experts to consider the ramifications and ethical hurdles of AI alongside its exciting opportunities.

Designers, policymakers and others share responsibility for ensuring safe and fair application of algorithms. The collaborative efforts of the AI Safety Institute Consortium are one of the leading examples of how industry leaders and stakeholders can pursue this goal.

"AI is accelerating the way any rapid and cheap writing tool is helpful to social processes of belonging. It creates shareable content that is made of familiar symbols and styles that seem to hail us an audience," Apple said.

Apple mentioned that a skepticism of such tools is wise precisely because they have so much potential to be abused. And according to Agnew, malicious actors are all too common, and automation could exacerbate ongoing practices that deny individuals control of their own representation, including doxxing, harassment, outing and deadnaming. The ongoing work of Queer in AI focuses on protecting LGBTQ+ people in all spaces, and most recently includes helping people who have transitioned maintain control over their identities in the academic and publishing worlds.

"Navigating the tensions of representation can be difficult, and sometimes that just explodes catastrophically when somebody loses control of their representation," Agnew said. "It has very dire consequences for them."

For Taylor, research into Reddit gave insight into the potential benefits of a more decentralized approach to applying algorithms, focusing on each community's needs and experience rather than a one-size-fits-all solution for online platforms.

"I think in general, it's oftentimes an overly simplistic binary to think about whether technology is good or bad. The question becomes, good and bad for whom, and in what contexts?" Taylor said.

Early insights from his study suggest that artists think about their work much differently than those who design or run platforms.

"One person said something about how the people who are most excited about generative AI in the context of art are CEOs who are looking to do the most work for the least amount of money, as opposed to the artists who—even if they want to make a living with their work—still find joy in the process of making.

"I think that joy, in its relationship to how people decide whether or not to use technology, is often undervalued," Taylor said.

More information: Rebuilding Trust: Queer in AI Approach to Artificial Intelligence Risk Management. docs.google.com/document/d/19d … ading=h.fkl2vs8ckg7z

Provided by Carnegie Mellon University

Citation: How should AI depict marginalized communities? Technologists look to a more inclusive future (2024, June 26) retrieved 29 June 2024 from https://techxplore.com/news/2024-06-ai-depict-marginalized-communities-technologists.html