FACTSHEET

AI + POSSIBLE REPRESENTATIONAL HARM FOR DIVERSE STUDENTS

Large language models (LLMs) – *artificial intelligence (AI) systems designed to understand, process, and produce human-like text* – generate outputs about people. These outputs are stories that can reinforce representational harms and stereotypes for diverse youth.

Faye Marie Vassel, a post-doctoral fellow at Stanford University's Institute for Human-Centered AI, describes research examining bias in large language models used in educational contexts to generate stories about students.* Vassel highlights the value of an intersectional lens in understanding how these AI depictions can impact young people.

Vassel describes three key forms of potential representational harm in AI outputs: **erasure**, **subordination**, and **stereotypes**.

01. ERASURE

The absence or near absence of certain groups.

Vassel describes the overwhelming lack of Indigenous stories in AI-generated outputs; when these stories do appear, Indigenous people are often positioned as *objects* of study rather than as learners themselves. For example, a student might be described as studying Indigenous history or traditional practices, while the Indigenous person is rarely represented as the learner. When Indigenous learners are depicted, the tool will sometimes use language that frames high achievement as unusual (such as "against all odds").

Representation affects visibility and belonging. When learners do not see themselves reflected, or only appear through limiting frames, AI outputs can contribute to their experience of marginalization in educational contexts.

02. SUBORDINATION

The positioning of some people as *less capable* or *less likely* to succeed.

Vassel gives the example of "Maria," described as the most highly feminized Latinx name in the study. But "Maria" is almost never cast as a star student, and in the rare cases in which she is, she is not shown in Science, Technology, Engineering & Mathematics (STEM) fields. This positioning reinforces the stereotype that Latinx learners need high levels of academic support and are not capable of being high performers in STEM contexts.

These kinds of portrayals matter because they shape ideas about who is 'allowed' to be seen as academically strong, and in which subjects.

03. STEREOTYPES

An oversimplified and generalized belief or assumption about a particular group of people.

Vassel shares the example of the 'model minority' stereotype, which describes a group of people (defined by race or ethnicity) who are seen as having achieved higher levels of success than other 'minority' groups. The large language models output stories in which "Priya" ("a typically Asian feminine name") was the star student. One model reinforced the stereotype that all Asians are "good" at STEM, presenting a monolithic view of Asian students.

Another example is the 'white savior' stereotype. In these stories, "Sarah" ("a name which has a high likelihood of being a white individual"), is depicted as academically strong across STEM, the humanities, and the social sciences. She is also often shown supporting students from identity groups that may be experiencing marginalization, such as "Jamal" and "Carlos", who are depicted as lacking agency.

These stereotypes can reinforce limitations on how diverse students see themselves and others.

IMPLICATIONS FOR EDUCATION

Representational harms can be especially damaging for students with intersecting identities. Depictions that **erase**, **subordinate**, and/or **stereotype** may affect students' sense of belonging and contribute to experiences of shame.

AI tools must be examined through a socio-technical and intersectional lens. Users' social identities must be centered in the development of large language models, and AI users must recognize that these tools are context-dependent and ensure that they are integrated in education in culturally-responsive ways.

---

*In ***How Harmful are AI's Biases on Diverse Student Populations?***, published on October 3, 2024.*