Employing Artificial Intelligence to Measure and Improve Representation in Children's Books

Anjali Adukia, The University of Chicago

Research shows that the educational resources instructors choose can greatly affect how children perceive their world. In particular, the way that people are represented within books can contribute to children's understanding about what social and professional roles they and others can or cannot inhabit. Given persistent racial and gender inequality in society and the importance of identity and representation in driving beliefs, aspirations, and academic effort and outcomes, these materials offer a key means to either address, perpetuate, or entrench core societal inequalities.

Given their impact, it is critical for those responsible for publishing and sharing these works to understand how their content stacks up. New software used for systematic analysis of images can be applied alongside established text analysis methods to analyze children's books, and gain a deeper understanding of what children currently have access to and what changes are needed going forward.

Tracking Diversity in Children's Books

Historically, identifying racial and gender bias in books has been done by human coders. This work has contributed to substantial progress around diversity, but it has also been a time-consuming effort, necessarily limited in scope and affected by human behavior and biases along the way.

My new study co-authored with Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz analyzed a series of books targeted to children and likely to appear in homes, classrooms, and school libraries over the past century. We developed a unique application of artificial intelligence tools that systematically identify and classify components of images at scale—in this case, detecting characters in photos and illustrations and classifying their race, gender, and age. While there are some biases in AI training data and algorithms, they can be easier to replicate and standardize, and can be applied to a much larger sample than manual content analysis. Using this tool, we were able to study 1,130 books to reveal trends of diversity in children's literature over time, as well as the current state of representation.

Opportunity to Do Better

Our AI tool was trained to detect faces, classify skin color, and predict the race, gender and age of the faces. We built on existing tools and made substantial improvements, including training the model to analyze illustrations, develop a classification of skin color, and introduce a higher precision for classification of gender and age. These precise measures were able to reveal more specific and sensitive trends that give us greater insight into the current state of representation in children's literature. Our analysis identified the following trends about race in children's books:

- Books have included more characters with dark skin over time
- Books classified as "mainstream" or written without the intent to highlight a specific identify group are more likely to depict lighter-skinned characters than those that highlight diversity or center underrepresented identities, even when they fall in the same racial category

- Black and Latinx people are underrepresented in the images and text, relative to their share of the U.S. population
- Across the collection, children were more likely than adults to be shown with lighter skin

We also looked closely at trends in female representation in children's literature, analyzing both female appearances in images and female mentions in text to find that:

- Female characters are more consistently seen in images than spoken about in the text, except in a collection of books specifically chosen to highlight female characters—suggesting that female characters are often included symbolically rather than substantively
- Male characters, especially white male characters, are persistently more likely to be represented, with little change over time
- Female characters appearing in images are predominately white

The ability of the AI software to analyze such a large subset of content opened the door to reveal these trends and new insights, a process that could be standardized and applied to greater analysis efforts.

Recommendations

Assessing representation in content traditionally involves the pain-staking time and effort of human coders analyzing each piece of information by hand. While these efforts have been made to improve the diverse representation in children's books, such efforts are slow and limited to smaller samples due to the labor-intensive nature of such a manual effort. AI software introduces a systematic, scalable approach to analyzing images and text in books and other visual content, which is complementary to traditional content analysis methods.

The software is not without its limitations; while this novel research program could allow for large-scale analysis to be done on the images and text in children's books, human coders are better able to capture the deeper nuances of content. By using these two methods in tandem, human coders would be more available to do in-depth text analysis, and researchers could provide a more comprehensive picture of the state of diversity in children's literature. The "optimal" level of representation varies by individual goals and purpose, and this work does not offer a specific prescription for authors or publishers. However, this tool offers the opportunity to better and more consistently measure representation, reveal continuing biases, and increase awareness by understanding what is in the "black box" of the content to which children are exposed.

For those invested in the continued effort to create better representation in children's books, tools such as those we developed can produce more and better evidence for their cause. Artificial intelligence can more easily and precisely inform publishers, content developers, and advocates of what work remains to be done, and help keep track of the changes that are made over time.