Ethics and Social Impact of AI Teachers

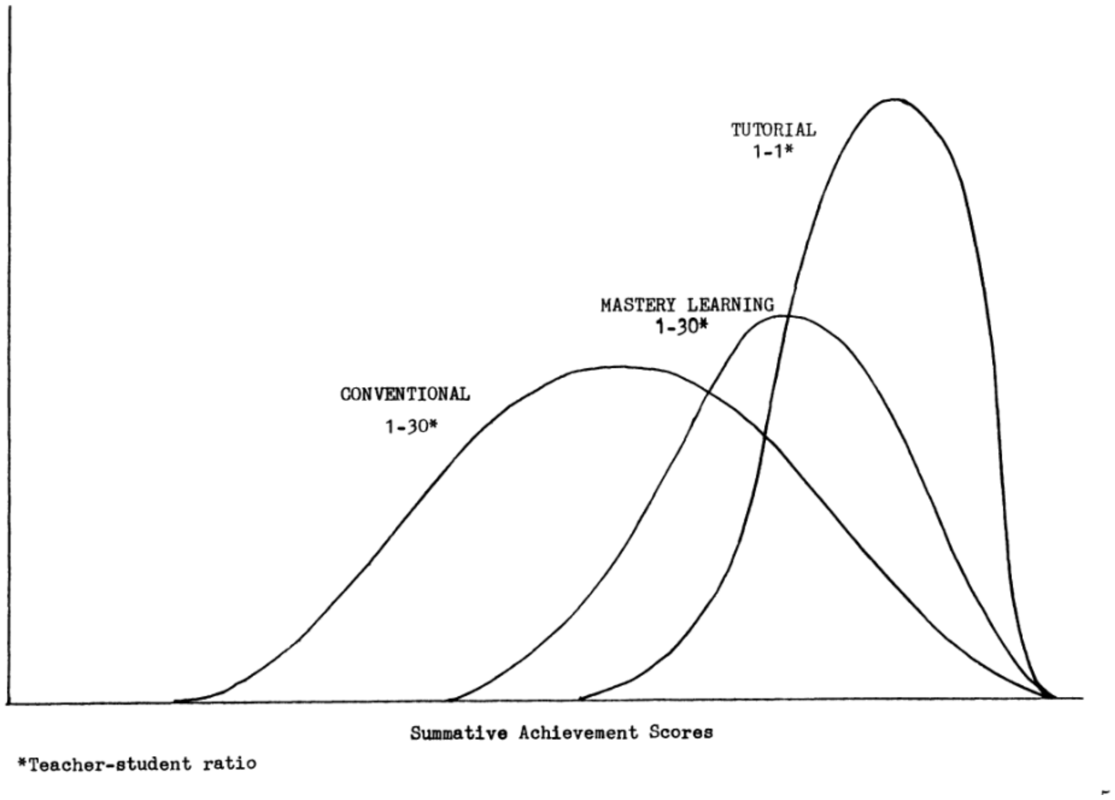
Qumeng Sun

26/06/2023

UNIVERSITÄT
GÖTTINGEN

The 2 Sigma Probem: The Search for Methods of Group Instrction as Effective as One-to-One Tutoring

FIGURE 1. *Achievement distribution for students under conventional, mastery learning, and tutorial instruction.*

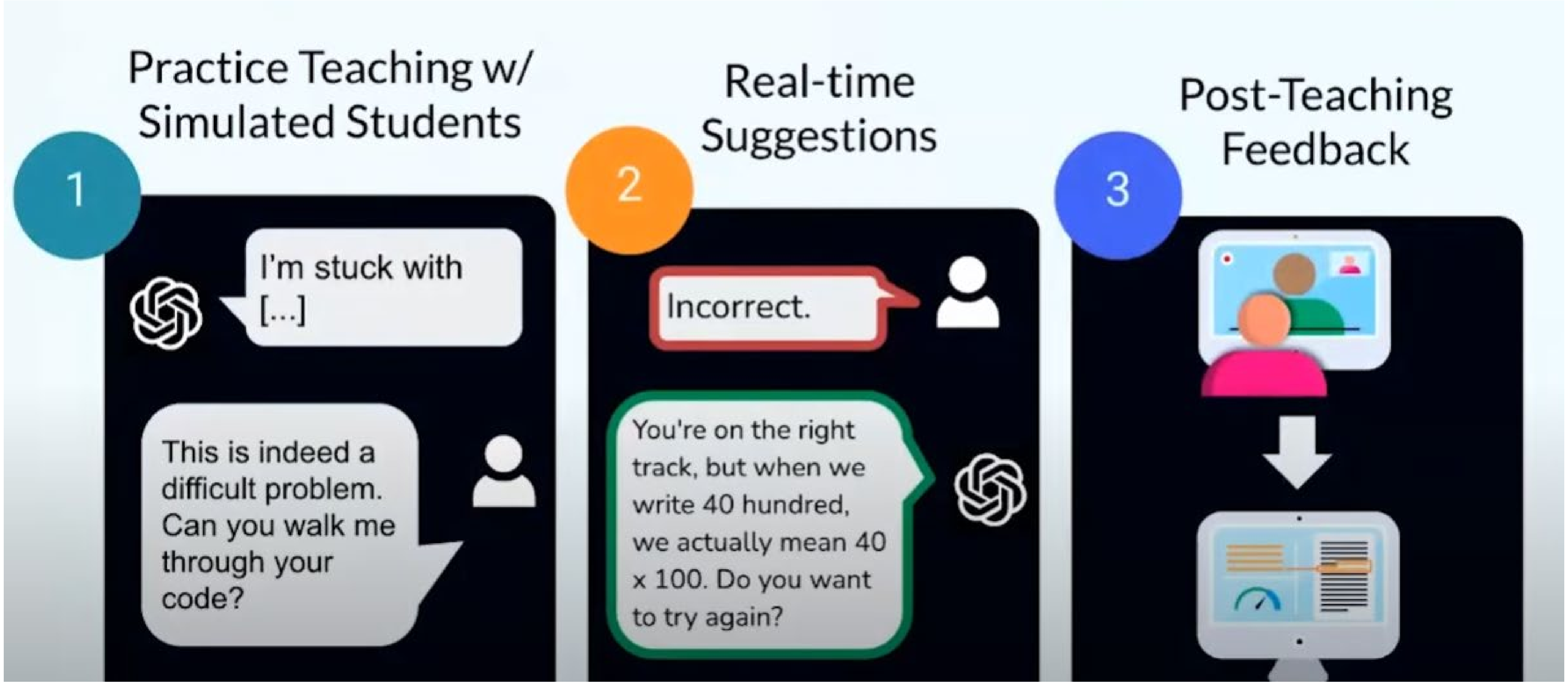


Students who learn using the TA model perform better and have better confidence.

This proves that one-on-one teaching assistants can be very effective in helping students get high marks on the exam quickly.

Bloom (1984)

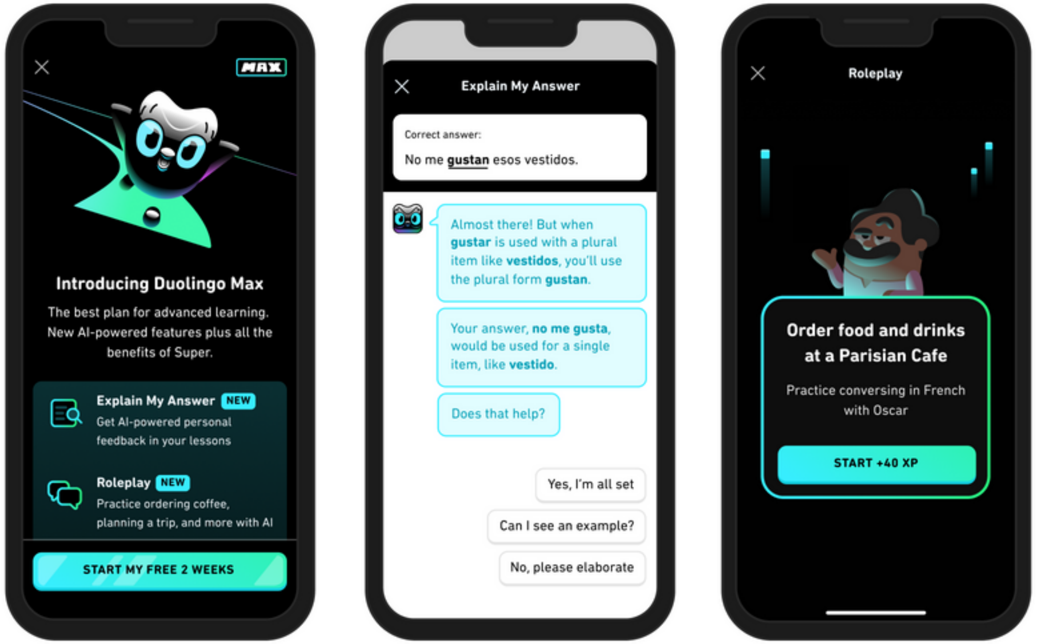
Khanmigo power by OpenAI



1
Practice Teaching w/ Simulated Students
I'm stuck with […]
This is indeed a difficult problem. Can you walk me through your code?
2
Real-time Suggestions
Incorrect.
You're on the right track, but when we write 40 hundred, we actually mean 40 x 100. Do you want to try again?
3
Post-Teaching Feedback

Source: [4]

Duolingo Max & Q-Chat



Introducing Duolingo Max

The best plan for advanced learning. New AI-powered features plus all the benefits of Super.

Explain My Answer NEW Get AI-powered personal feedback in your lessons

Roleplay NEW Practice ordering coffee, planning a trip, and more with AI

Correct answer:
No me gustan esos vestidos.

Almost there! But when gustar is used with a plural item like vestidos, you'll use the plural form gustan.

Your answer, no me gusta, would be used for a single item, like vestido.

Does that help?

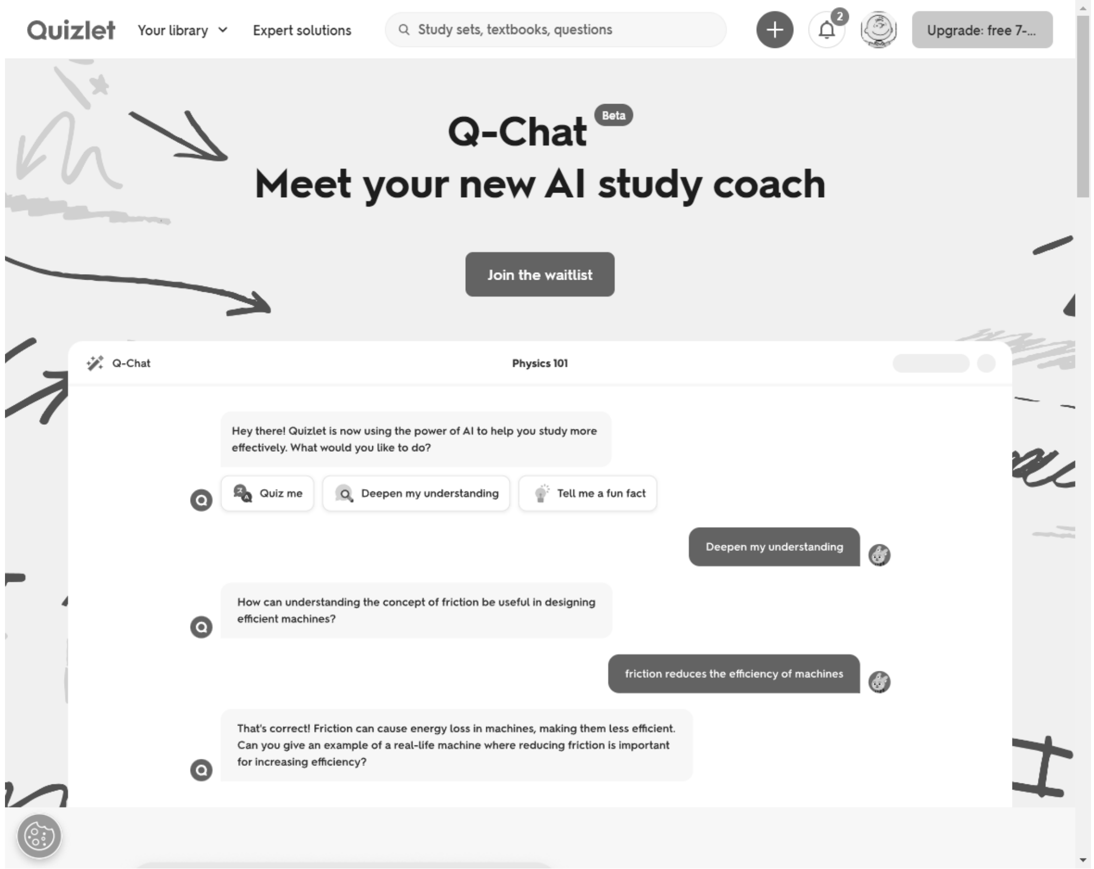
Yes, I'm all set

Can I see an example?

No, please elaborate

Order food and drinks at a Parisian Cafe

Practice conversing in French with Oscar



Q-Chat Beta

Meet your new AI study coach

Hey there! Quizlet is now using the power of AI to help you study more effectively. What would you like to do?

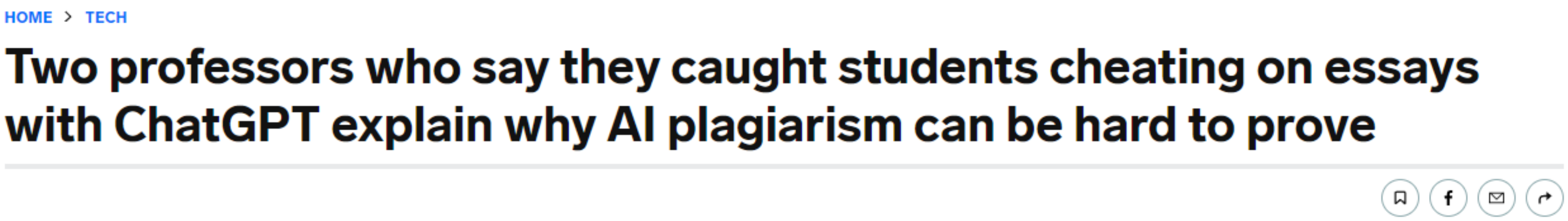
Deepen my understanding

How can understanding the concept of friction be useful in designing efficient machines?

friction reduces the efficiency of machines

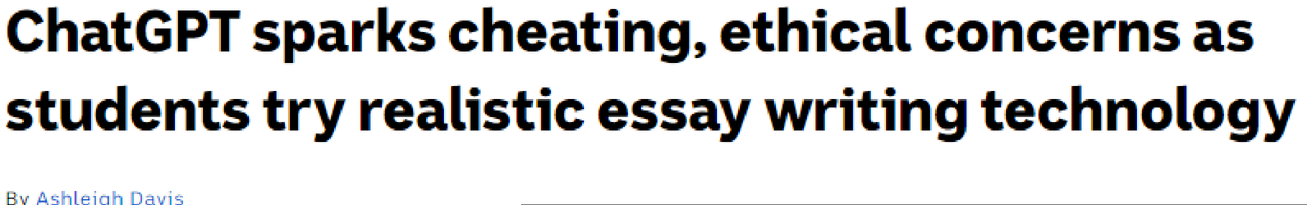
That's correct! Friction can cause energy loss in machines, making them less efficient. Can you give an example of a real-life machine where reducing friction is important for increasing efficiency?

Source: Introducing Duolingo Max, a learning experience powered by GPT-4
https://blog.duolingo.com/duolingo-max/
Q-Chat Beta Meet your new AI study coach
https://quizlet.com/labs/qchat



HOME > TECH

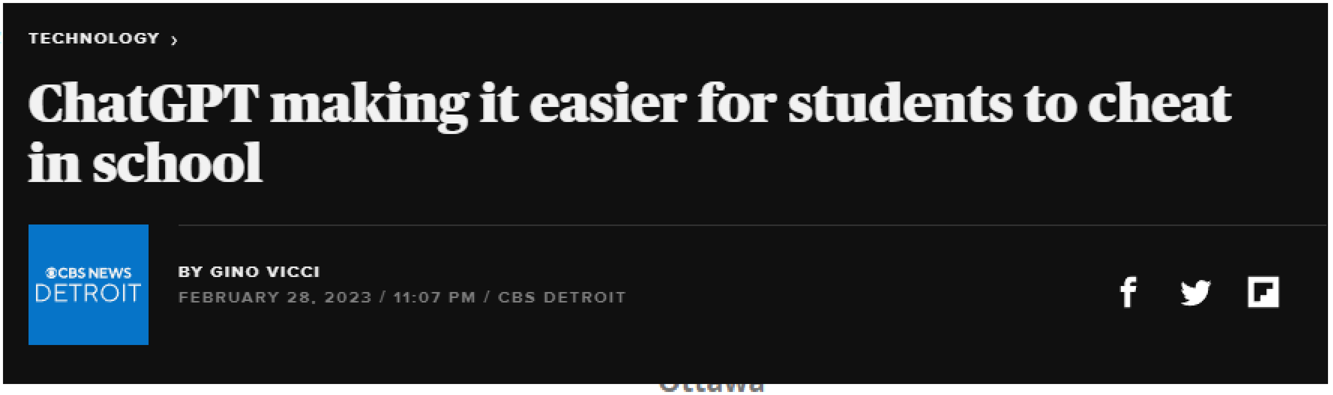
Two professors who say they caught students cheating on essays with ChatGPT explain why AI plagiarism can be hard to prove



ChatGPT sparks cheating, ethical concerns as students try realistic essay writing technology

By Ashleigh Davis

Posted Wed 25 Jan 2023 at 11:16pm, updated Thu 2



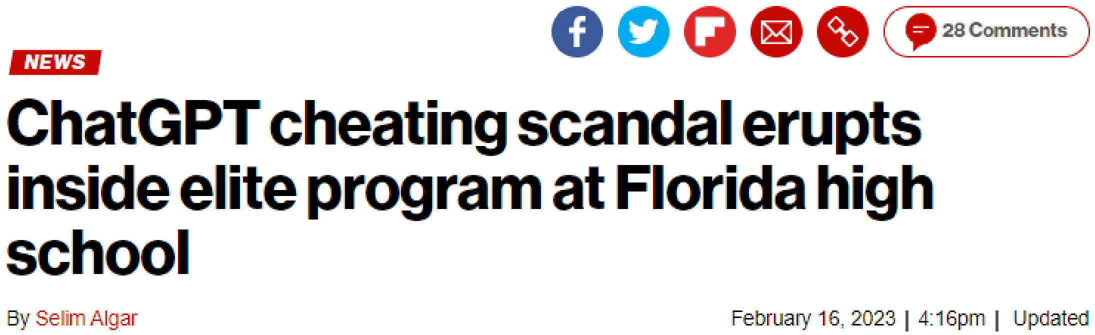
TECHNOLOGY ›

ChatGPT making it easier for students to cheat in school

CBS NEWS DETROIT

BY GINO VICCI

FEBRUARY 28, 2023 / 11:07 PM / CBS DETROIT



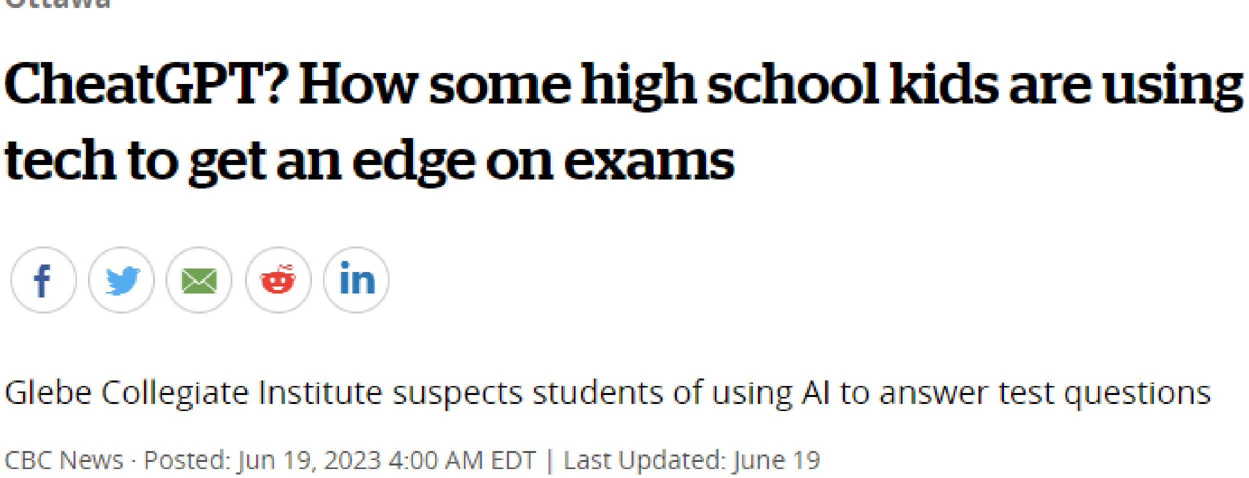
28 Comments

NEWS

ChatGPT cheating scandal erupts inside elite program at Florida high school

By Selim Algar

February 16, 2023 | 4:16pm | Updated



Ottawa

CheatGPT? How some high school kids are using tech to get an edge on exams

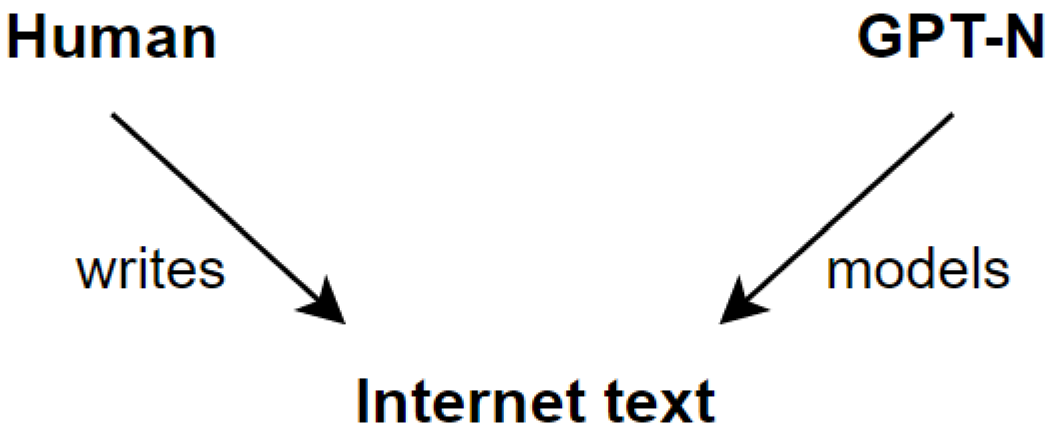
Glebe Collegiate Institute suspects students of using AI to answer test questions

CBC News · Posted: Jun 19, 2023 4:00 AM EDT | Last Updated: June 19

Ethics Within the Model

1. Truthfulness (misleading) [4]
2. Privacy protection [4]
3. Harmful Information

1. Truthfulness



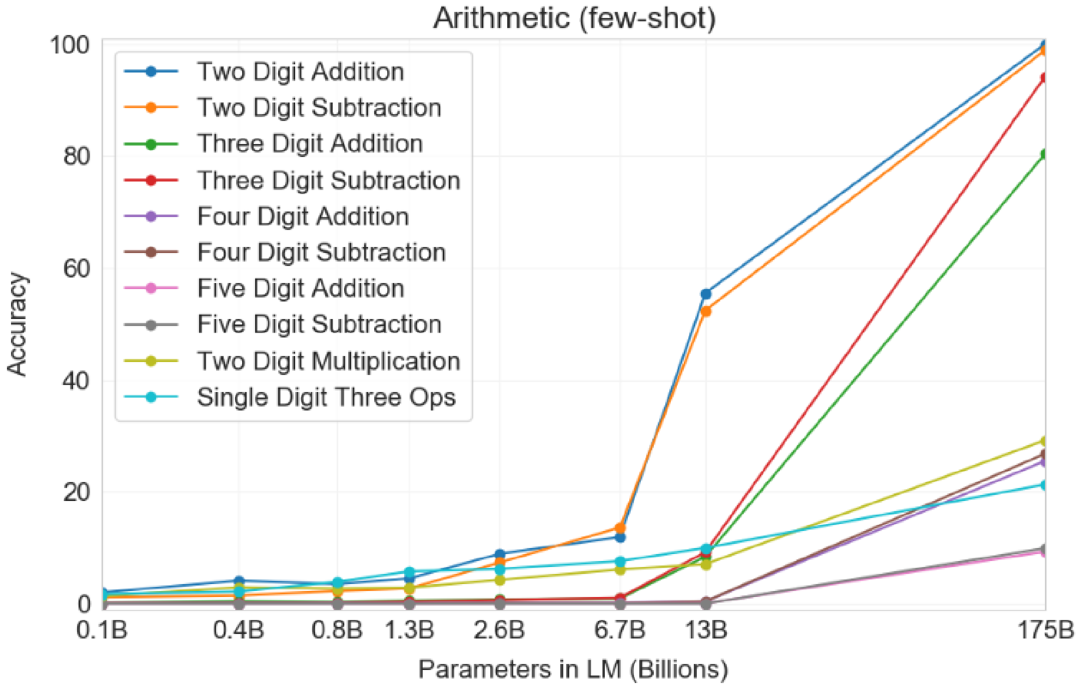
Human

writes

GPT-N

models

Internet text

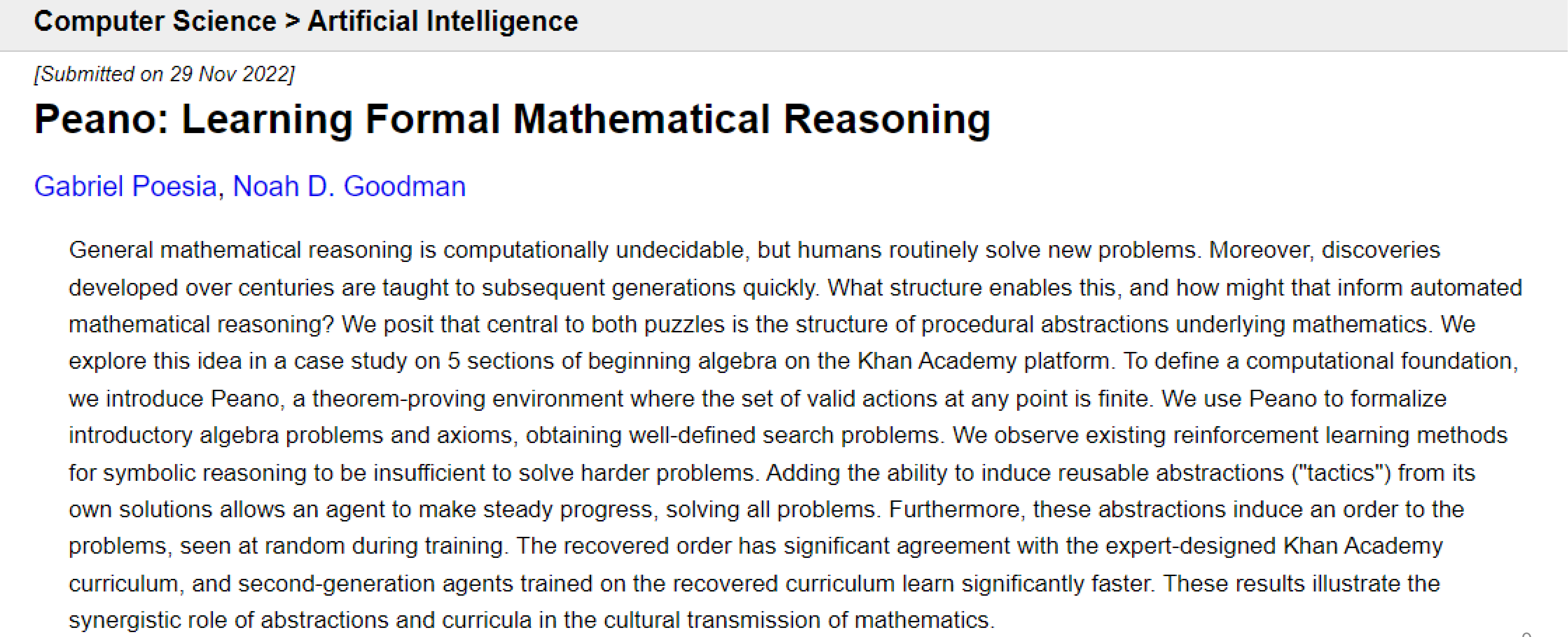


AI can lie. They also bad at doing math

Source: Is GPT-N bounded by human capabilities? No.
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.lesswrong.com%2Fposts%2FbZbLnr7qwuEBpTPuF%2Fis-gpt-n-bounded-by-human-capabilities-no&psig=AOvVaw09zDT8sXPixcqSCN_zFZOj&ust=1687729937750000&source=images&cd=vfe&ved=0CBMQjhxqFwoTCMiT8ZDy3P8CFQAAAAAdAAAAABAE

Truthfulness – Solutions

[7] Effectively improve the weakness of AI for mathematical problems.



Computer Science > Artificial Intelligence

[Submitted on 29 Nov 2022]

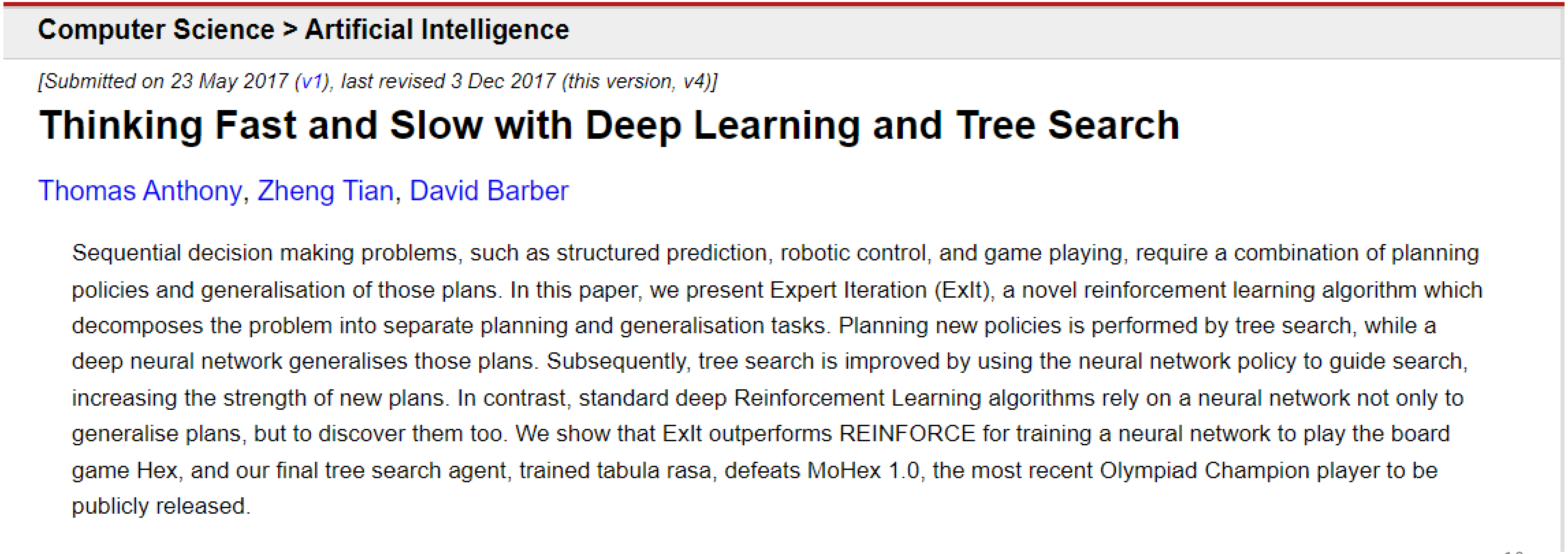
Peano: Learning Formal Mathematical Reasoning

Gabriel Poesia, Noah D. Goodman

General mathematical reasoning is computationally undecidable, but humans routinely solve new problems. Moreover, discoveries developed over centuries are taught to subsequent generations quickly. What structure enables this, and how might that inform automated mathematical reasoning? We posit that central to both puzzles is the structure of procedural abstractions underlying mathematics. We explore this idea in a case study on 5 sections of beginning algebra on the Khan Academy platform. To define a computational foundation, we introduce Peano, a theorem-proving environment where the set of valid actions at any point is finite. We use Peano to formalize introductory algebra problems and axioms, obtaining well-defined search problems. We observe existing reinforcement learning methods for symbolic reasoning to be insufficient to solve harder problems. Adding the ability to induce reusable abstractions ("tactics") from its own solutions allows an agent to make steady progress, solving all problems. Furthermore, these abstractions induce an order to the problems, seen at random during training. The recovered order has significant agreement with the expert-designed Khan Academy curriculum, and second-generation agents trained on the recovered curriculum learn significantly faster. These results illustrate the synergistic role of abstractions and curricula in the cultural transmission of mathematics.

Truthfulness – Solutions

[8] Constrain the next steps to apply the axioms efficiently.



Computer Science > Artificial Intelligence

[Submitted on 23 May 2017 (v1), last revised 3 Dec 2017 (this version, v4)]

Thinking Fast and Slow with Deep Learning and Tree Search

Thomas Anthony, Zheng Tian, David Barber

Sequential decision making problems, such as structured prediction, robotic control, and game playing, require a combination of planning policies and generalisation of those plans. In this paper, we present Expert Iteration (ExIt), a novel reinforcement learning algorithm which decomposes the problem into separate planning and generalisation tasks. Planning new policies is performed by tree search, while a deep neural network generalises those plans. Subsequently, tree search is improved by using the neural network policy to guide search, increasing the strength of new plans. In contrast, standard deep Reinforcement Learning algorithms rely on a neural network not only to generalise plans, but to discover them too. We show that ExIt outperforms REINFORCE for training a neural network to play the board game Hex, and our final tree search agent, trained tabula rasa, defeats MoHex 1.0, the most recent Olympiad Champion player to be publicly released.

Truthfulness Problems can be solved easily

w/ Reinforcement learning/Searching Strategies/Fine-tuning…

2. Privacy protection

1. Information collection should be It should be collected in a fair way, ideally with the knowledge or consent of the person whose data it is.
2. Information needs to be accurate, complete, and up-to-date.
3. Information should only be collected for a specific purpose that is clearly specified before collection The information should not be reused later in ways that are incompatible with the original specific purpose.
4. Information should not be shared or reused in ways incompatible with the original specific purpose without either consent of the person whose data it is or a law providing the authority to do so.
5. Information should be protected through reasonable security safeguards that limit the risk of an unauthorised person accessing, using, changing, sharing, or destroying that information.
6. There should be transparency about how information is collected, used, shared and protected.
7. People have certain rights about their own information, including the right to know what information others have about them, who has it, and the right to request that information be corrected, amended or erased.
8. Those holding data must be held accountable for the above principles.

Teaching Privacy and Ethical Guardrails for
The AI Imperative in Education
https://publicinterestprivacy.org/teaching-privacy-and-ethical-guardrails-for-the-ai-imperative-in-education/

Privacy protection

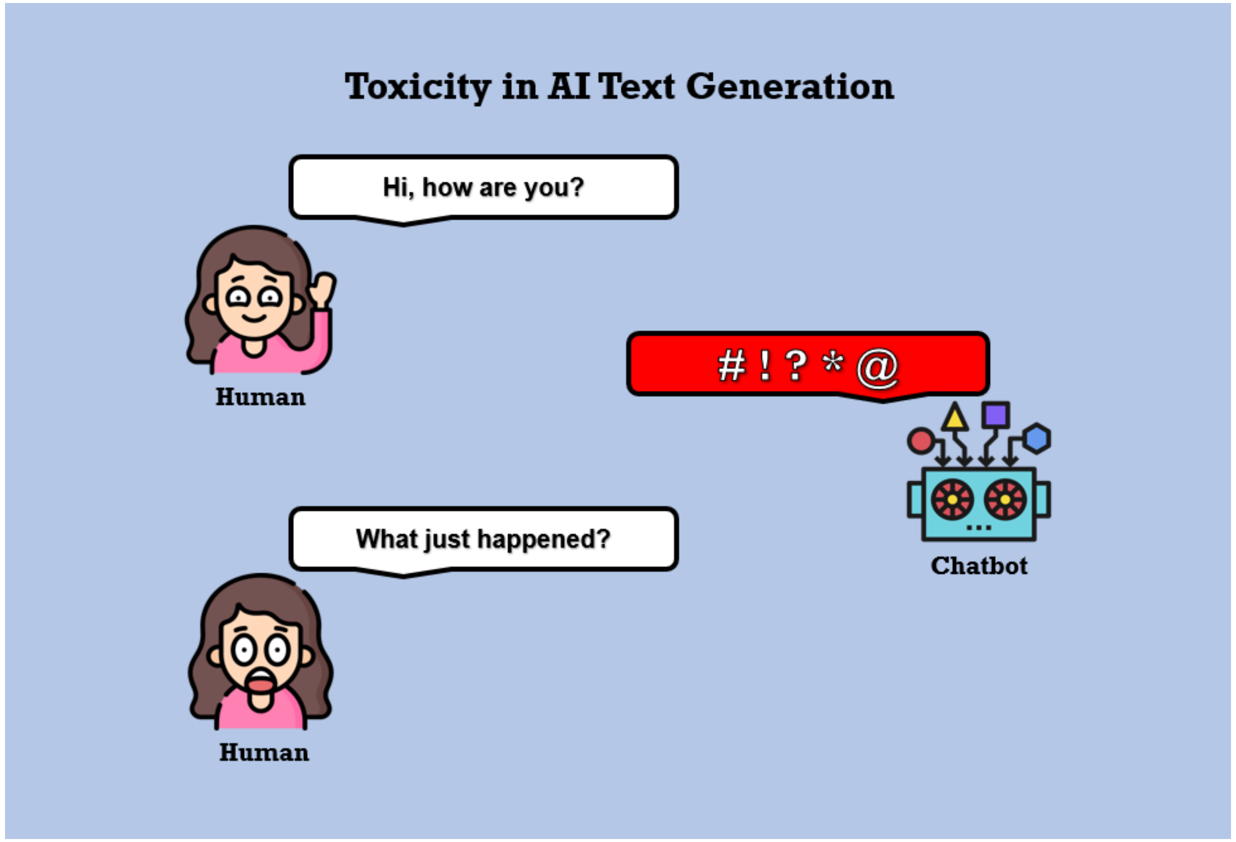
1. De-sensitive information[4]
2. Privately owned AI[4]

We can help students without harming their privacy.

Enabling AI-driven health advances without sacrificing patient privacy
https://news.mit.edu/2021/secure-ai-labs-health-care-1007

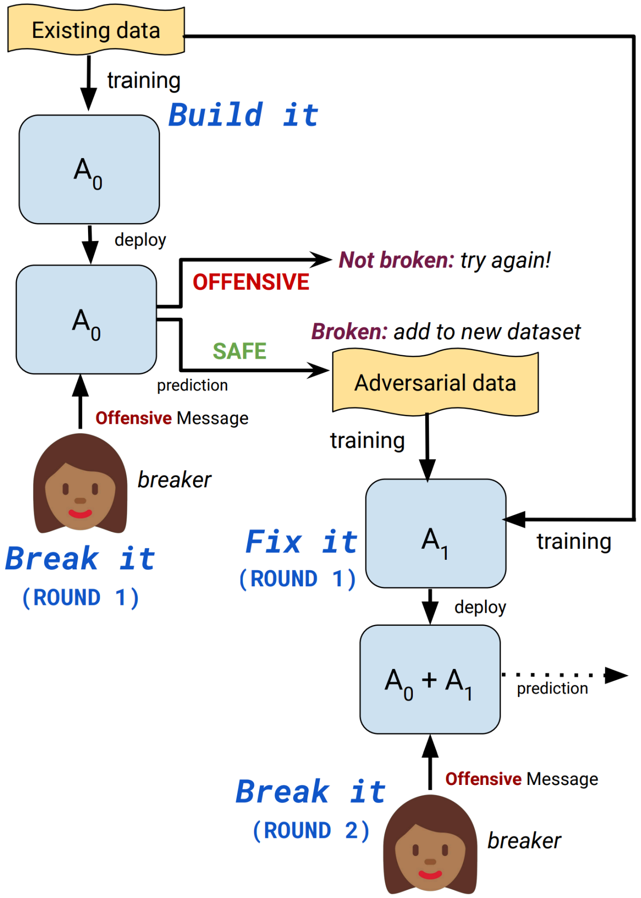
3. Harmful Information

- Toxic
- Hate speech
- Abusive language
- Anti-social speech
  - “How to make a bomb?”
- Bias & Discrimination
  - Racism, sexism, stereotypical influences, etc.
- …



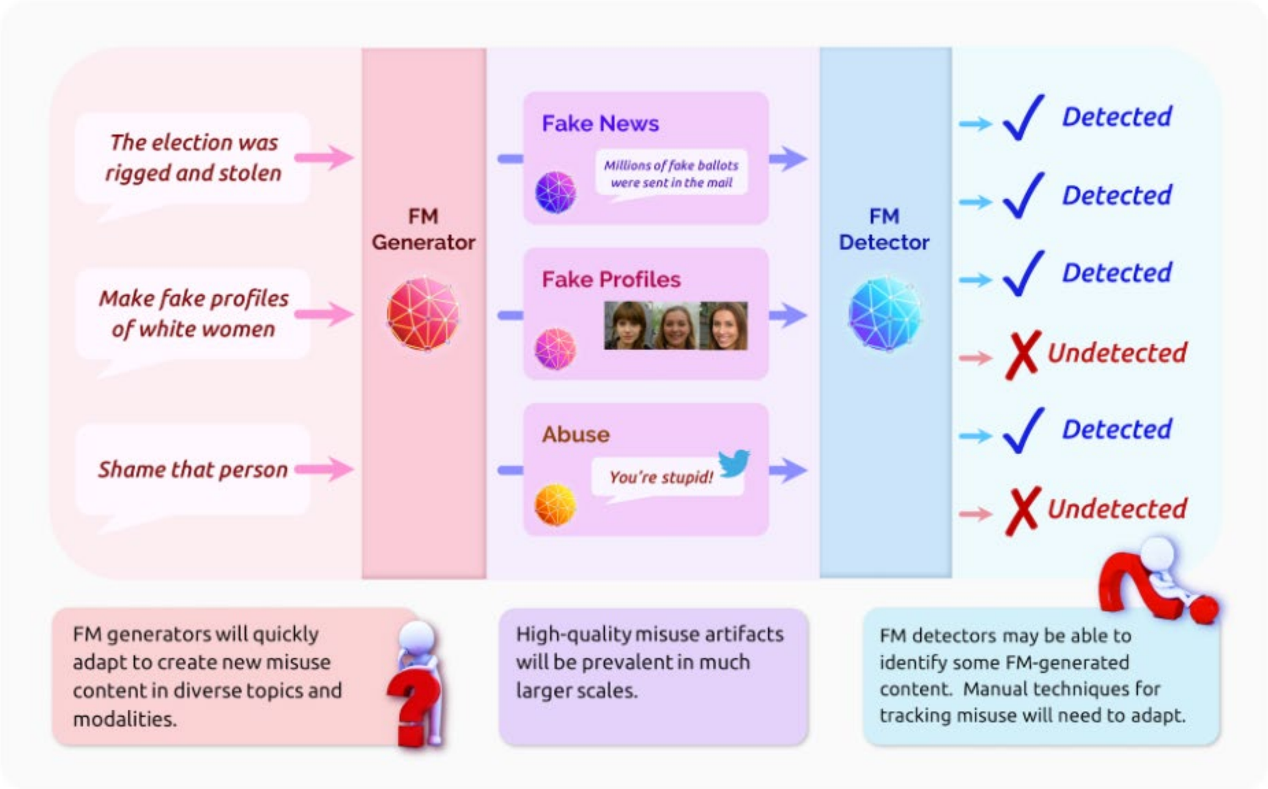
Graphic created by Julia Nikulski illustrating toxic language generation by a chatbot in a conversation with a user. Icons made by Becris and Freepik from Flaticon.

3. Harmful Information - Solutions



Source: Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack

FM Detector



FM generators will quickly adapt to create new misuse content in diverse topics and modalities.

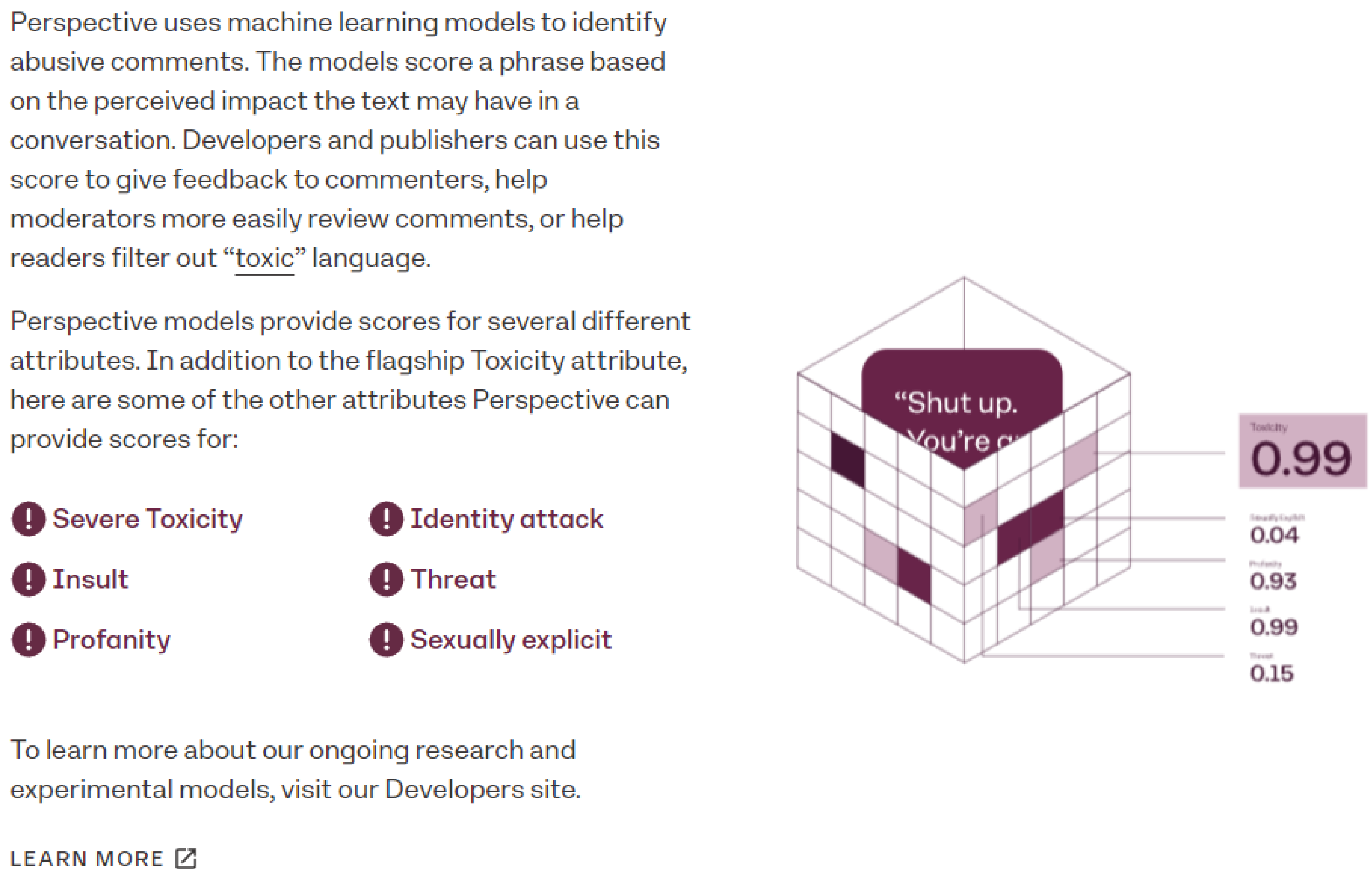
High-quality misuse artifacts will be prevalent in much larger scales.

FM detectors may be able to identify some FM-generated content. Manual techniques for tracking misuse will need to adapt.

Fig. 25. This figure shows the effect foundation models will have on manipulative and harmful content generation, and the implications for detection.

Source [6]

Perspective API



Perspective uses machine learning models to identify abusive comments. The models score a phrase based on the perceived impact the text may have in a conversation. Developers and publishers can use this score to give feedback to commenters, help moderators more easily review comments, or help readers filter out "toxic" language.

Perspective models provide scores for several different attributes. In addition to the flagship Toxicity attribute, here are some of the other attributes Perspective can provide scores for:

- Severe Toxicity
- Insult
- Profanity
- Identity attack
- Threat
- Sexually explicit

To learn more about our ongoing research and experimental models, visit our Developers site.

LEARN MORE

https://www.perspectiveapi.com/how-it-works/

3. Harmful - Compared to humans

- **Unable to pursue responsibility**
- **AI is more easily controlled and supervised**
  - Rules can limit AI, and its uncertainty is low. We cannot predict any behavior of a human teacher with 100% acc.
- **AI poses less of a threat to students.**
  - AI cannot physically harm students.
  - The AI has no self-interest, so there is no incentive to harm the student.

Ethics Outside the Model

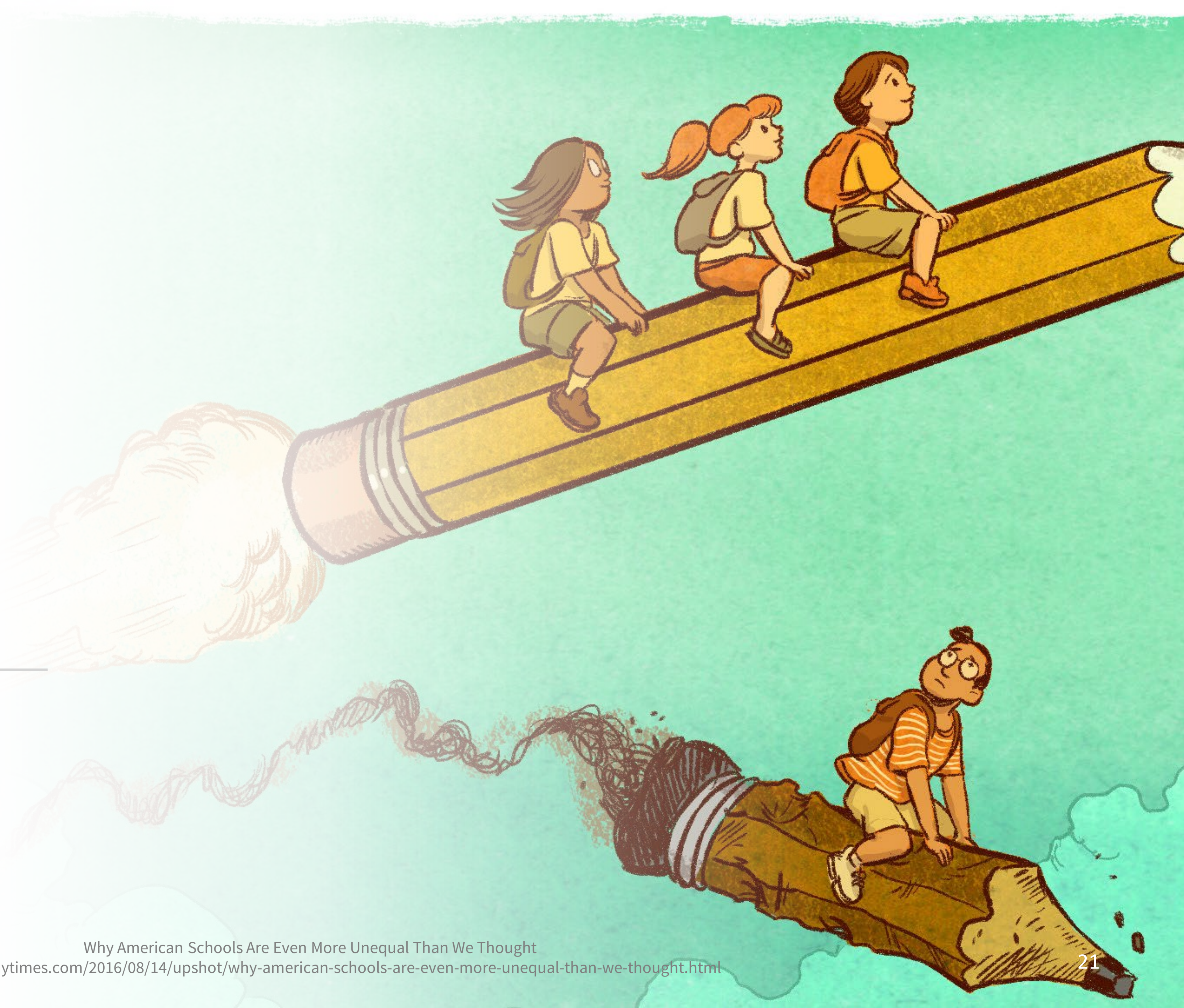
Fairness & Justice

Bias

Environment

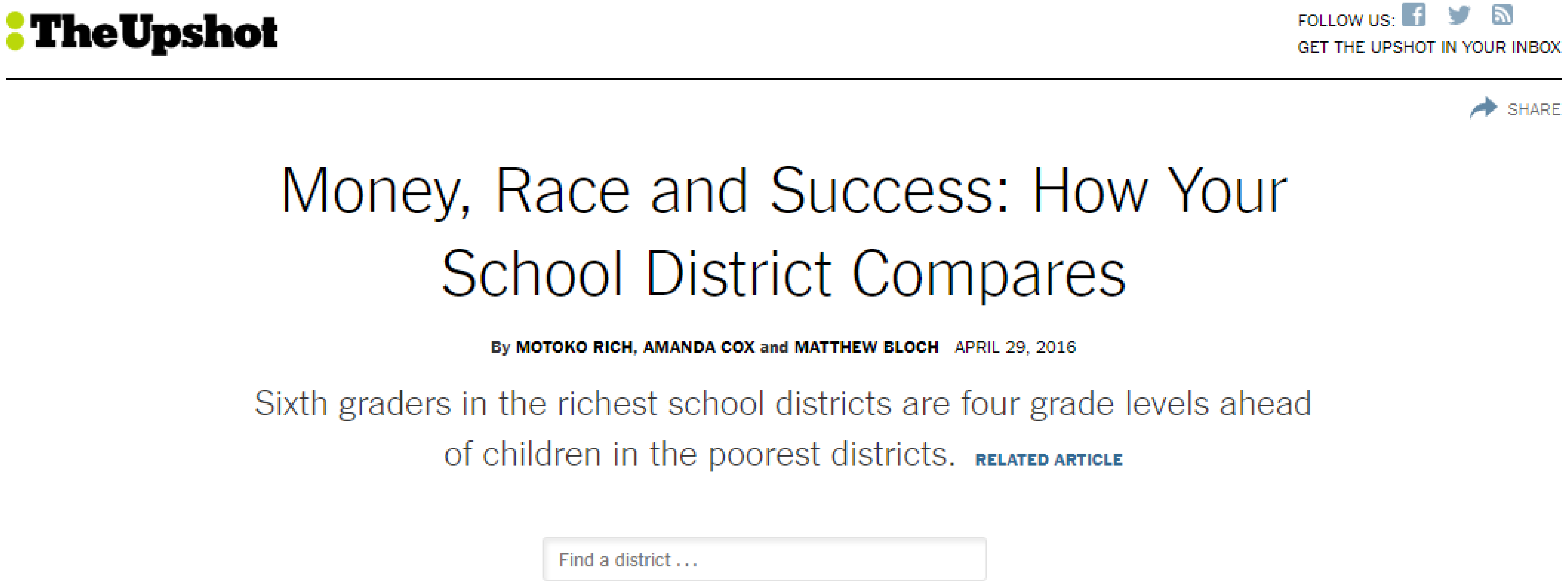
1. Fairness & Justice

Educational Equity



Why American Schools Are Even More Unequal Than We Thought
https://www.nytimes.com/2016/08/14/upshot/why-american-schools-are-even-more-unequal-than-we-thought.html

Even in the United States, unequal distribution of education is still a problem



Money, Race and Success: How Your School District Compares

By MOTOKO RICH, AMANDA COX and MATTHEW BLOCH APRIL 29, 2016

Sixth graders in the richest school districts are four grade levels ahead of children in the poorest districts. RELATED ARTICLE

Money, Race and Success: How Your School District Compares
https://www.nytimes.com/interactive/2016/04/29/upshot/money-race-and-success-how-your-school-district-compares.html?_r=0&mtrref=undefined&assetType=PAYWALL&mtrref=www.nytimes.com&gwh=A37970F232ECA63F157FF6FEB89B8A8D&gwt=pay&assetType=PAYWALL

Educational Equity

- Providing affordable education, Improving the overall quality of education in society, especially for vulnerable groups

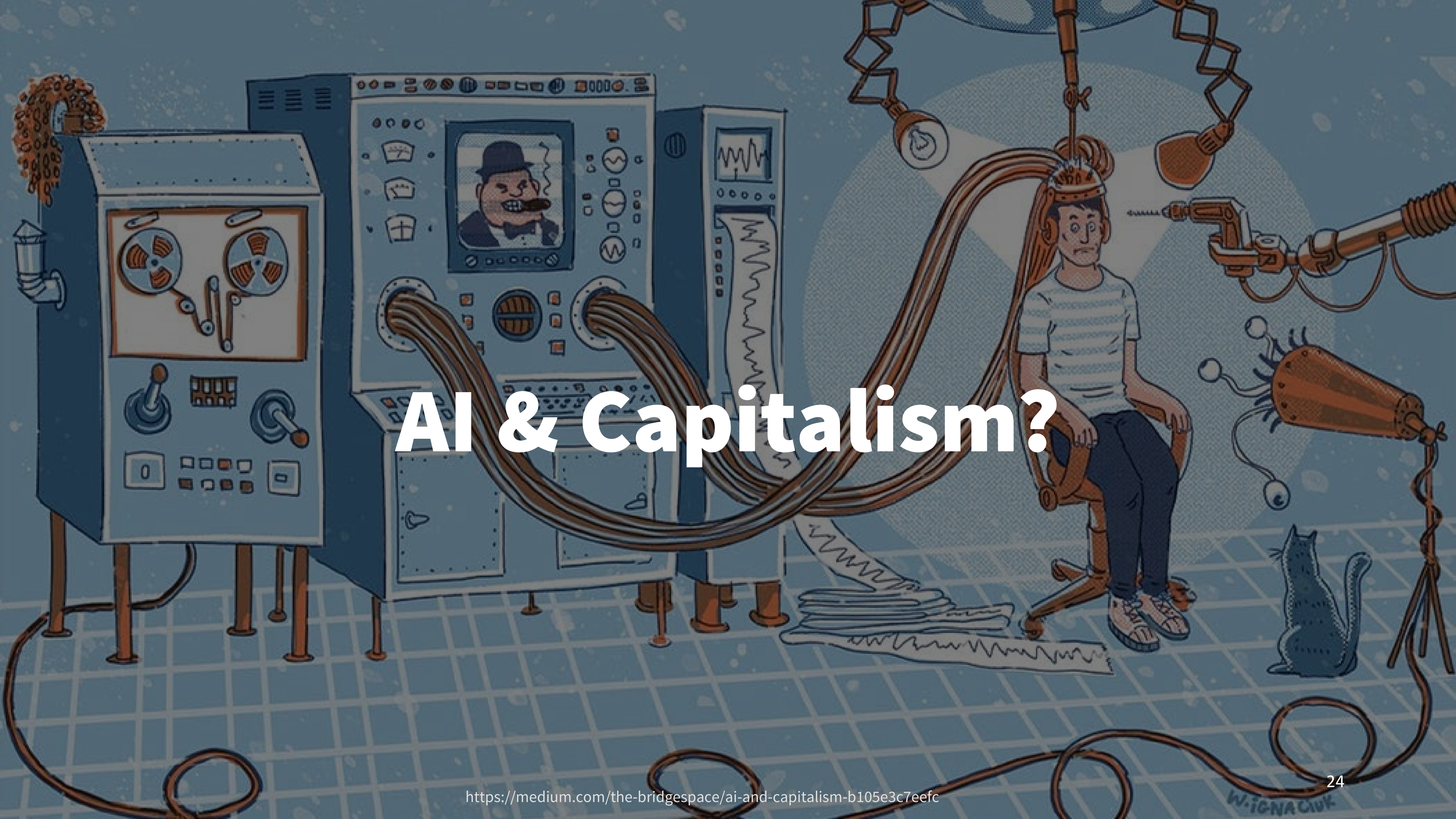
Washington, D.C., US

https://csipresident.fandom.com/wiki/Sidwell_Friends_School

Hunan, China

http://gs.cnr.cn/gsxw/tpxw/200709/t20070910_504564887.html

AI & Capitalism?

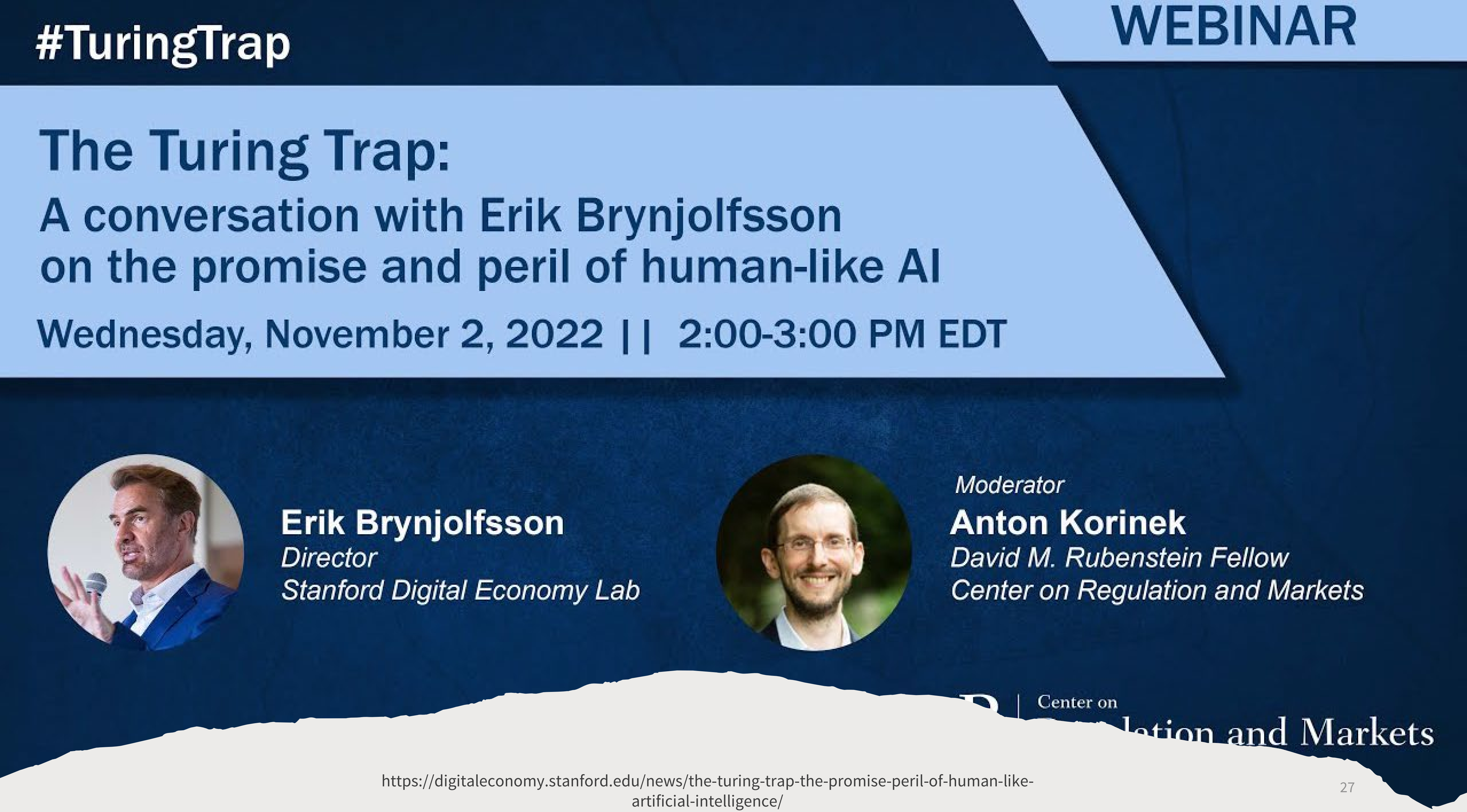


https://medium.com/the-bridgespace/ai-and-capitalism-b105e3c7eefc

GPT-5/6/...

GPT-X

https://www.youtube.com/watch?v=4joMUrqOWO8



#TuringTrap

WEBINAR

The Turing Trap:

A conversation with Erik Brynjolfsson on the promise and peril of human-like AI

Wednesday, November 2, 2022 || 2:00-3:00 PM EDT

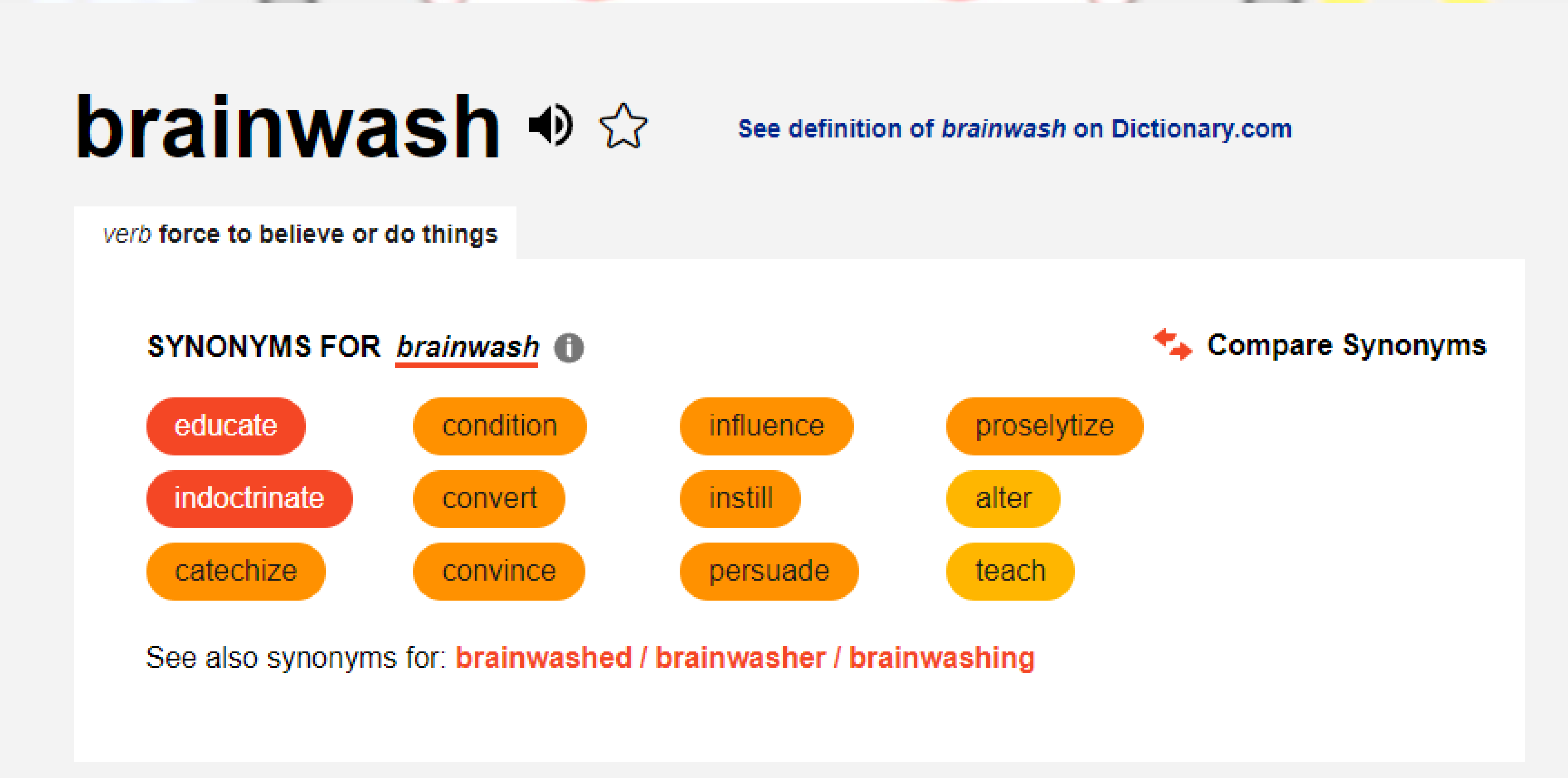
Erik Brynjolfsson
Director
Stanford Digital Economy Lab

Moderator
Anton Korinek
David M. Rubenstein Fellow
Center on Regulation and Markets

Center on
lation and Markets

https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-artificial-intelligence/

Misguided Values



brainwash

See definition of *brainwash* on Dictionary.com

verb force to believe or do things

SYNONYMS FOR *brainwash*

Compare Synonyms

educate · condition · influence · proselytize

indoctrinate · convert · instill · alter

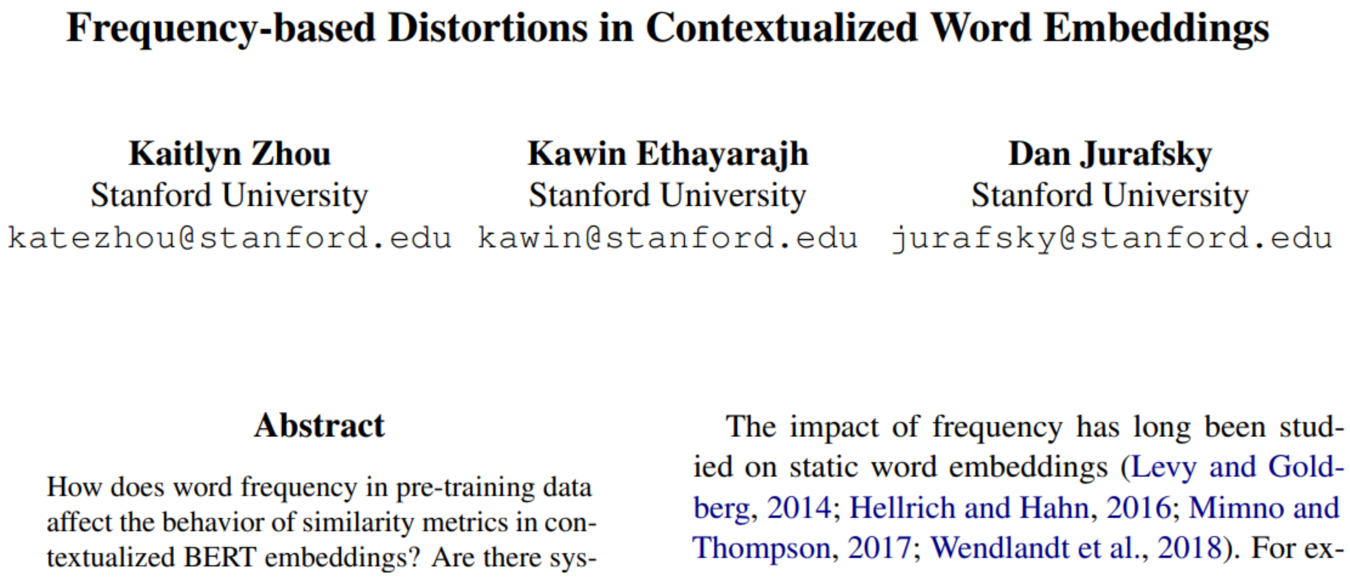
catechize · convince · persuade · teach

See also synonyms for: brainwashed / brainwasher / brainwashing

This is a Call-to-Arms Against AI-Generated Content
https://www.google.com/url?sa=i&url=https%3A%2F%2Feitanblumin.com%2F2023%2F05%2F27%2Fthis-is-a-call-to-arms-against-ai-generated-content%2F&psig=AOvVaw1mAkSaagVqlo5HQe8sofOg&ust=1687792281351000&source=images&cd=vfe&ved=0CBMQjhxqFwoTCJjd9rDa3v8CFQAAAAAdAAAAABAK
https://www.thesaurus.com/browse/brainwash

Anglocentric perspective

- BERT-Base has more trouble differentiating between South American and African countries than North American and European ones.



Frequency-based Distortions in Contextualized Word Embeddings

Kaitlyn Zhou
Stanford University
katezhou@stanford.edu

Kawin Ethayarajh
Stanford University
kawin@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

How does word frequency in pre-training data affect the behavior of similarity metrics in contextualized BERT embeddings? Are there sys-

The impact of frequency has long been studied on static word embeddings (Levy and Goldberg, 2014; Hellrich and Hahn, 2016; Mimno and Thompson, 2017; Wendlandt et al., 2018). For ex-

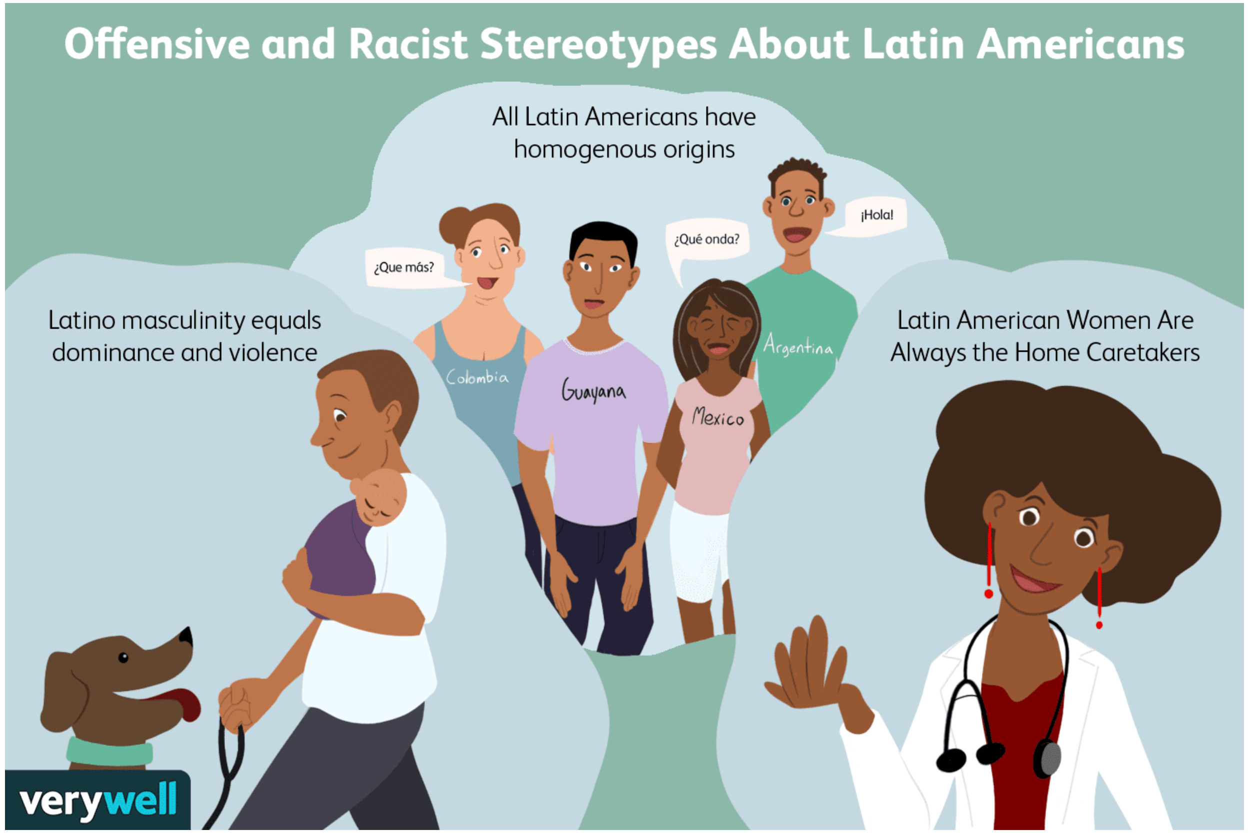
“This imbalance of word and topic frequency ultimately reflects an Anglocentric and Eurocentric view of the world.” [16]

1. Fairness & Justice - Solutions

- Ensure neutrality and **diversity** of sources.
- **“Magic Must Defeat Magic,”** AI’s detector for Harmful Information.
- Increasing the **transparency** of research at large companies to the public
- The **government** strictly monitors AI teachers' research and use by enacting **new laws**.

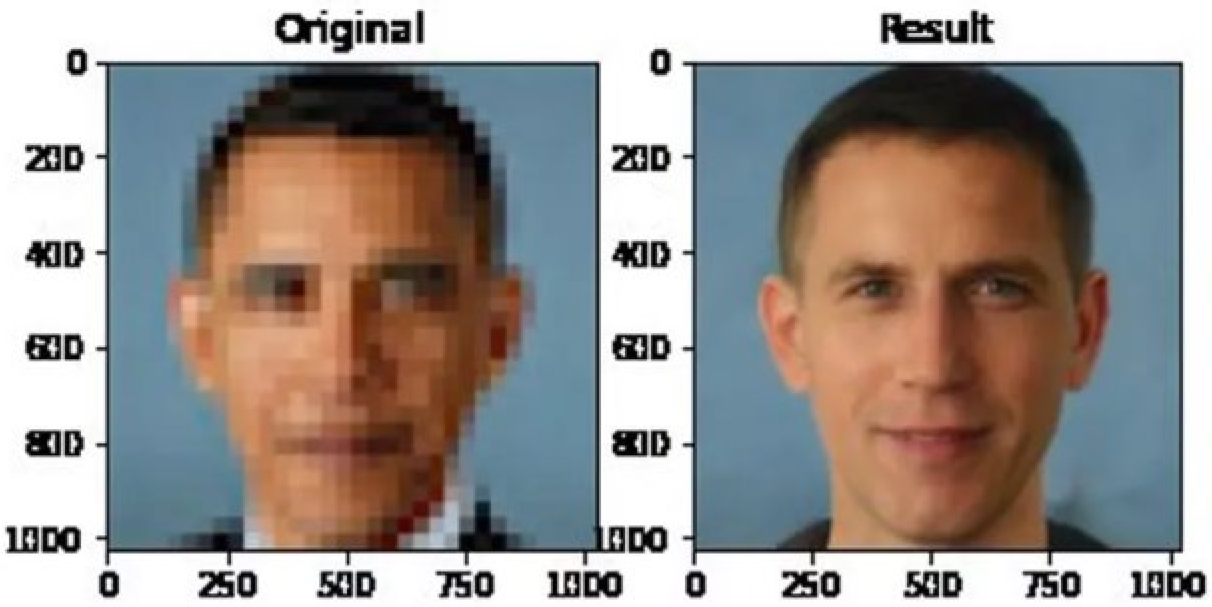
2. (Representational) Bias

- Harmful stereotypes
- Underrepresentation
- Overrepresentation

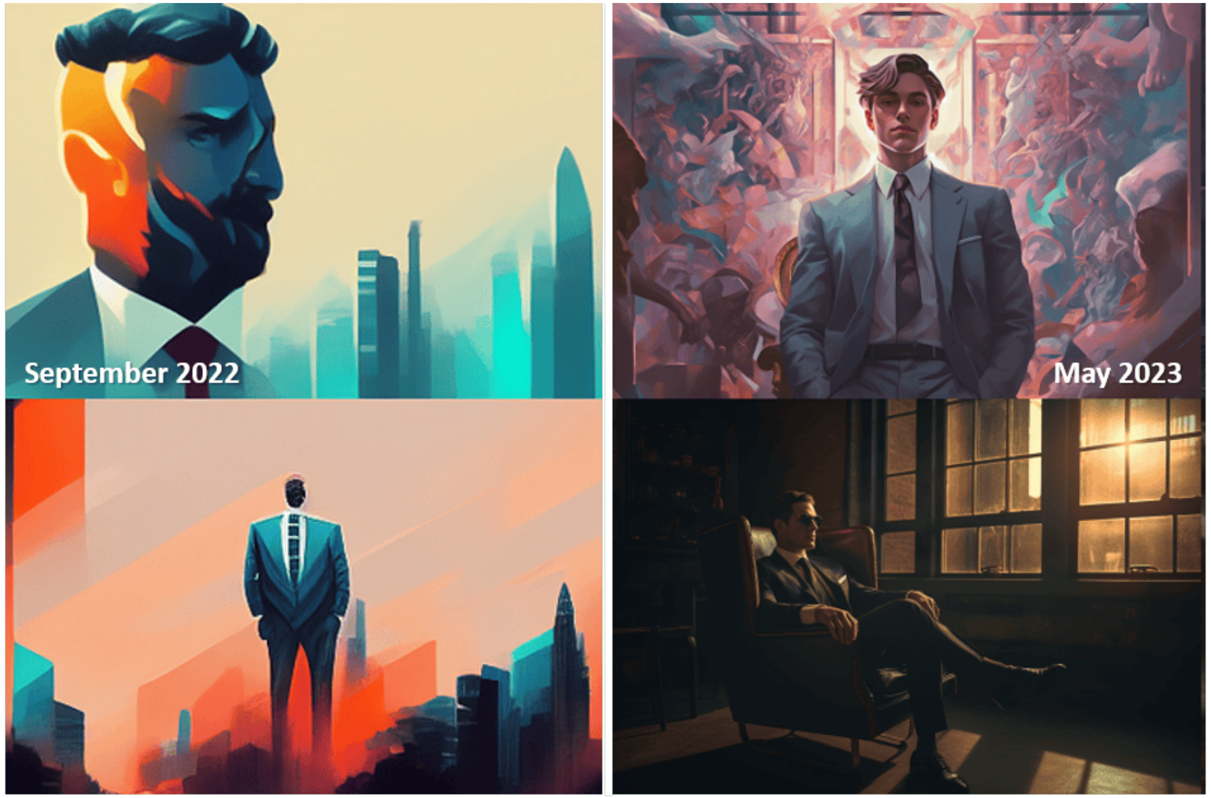

Offensive and Racist Stereotypes About Latin Americans
All Latin Americans have homogenous origins
¿Que más?
¿Qué onda?
¡Hola!
Colombia
Guayana
Mexico
Argentina
Latino masculinity equals dominance and violence
Latin American Women Are Always the Home Caretakers
verywell

6 Harmful Stereotypes About Latin Americans
https://www.verywellmind.com/6-harmful-stereotypes-about-latin-americans-5113358

Bias is everywhere in AI industry



Bias in Computer Vision

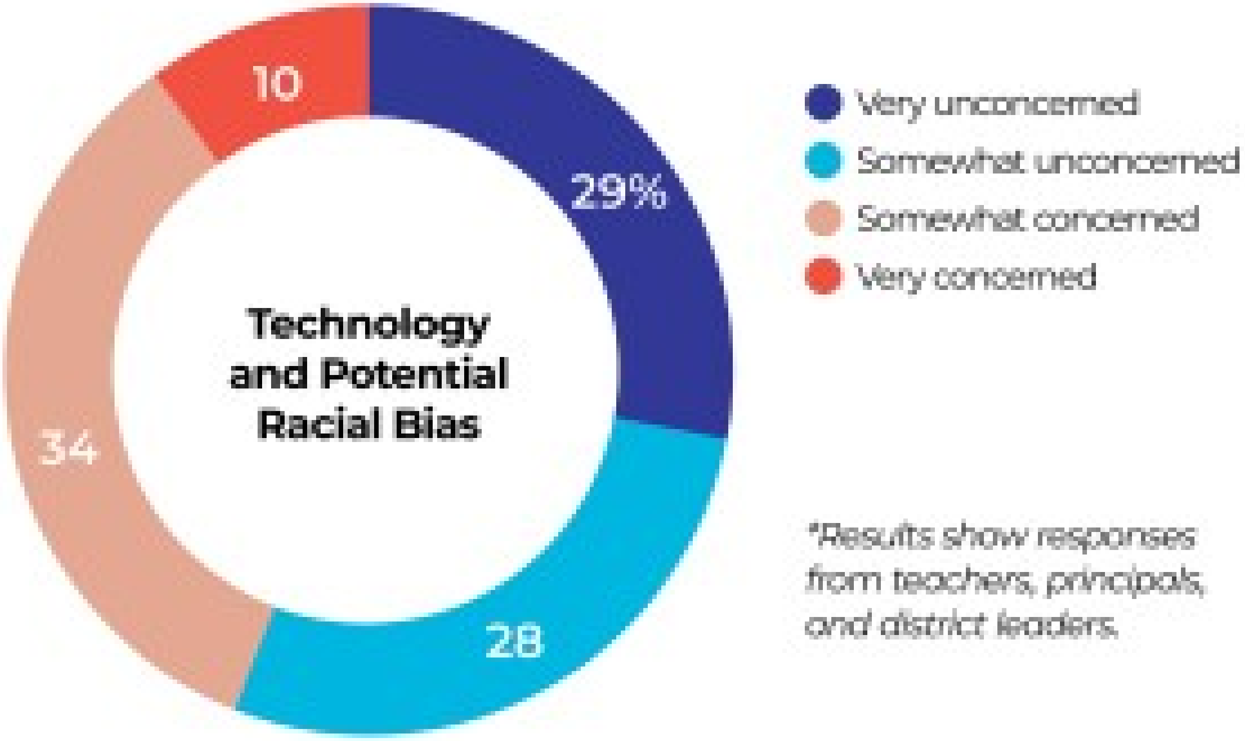


CEO, generated by AI

https://adagetechnologies.com/ai-discrimination/

https://www.wearedevelopers.com/magazine/generative-ai-image-bias

How concerned are you about the potential for racial bias in the educational technology used in schools?



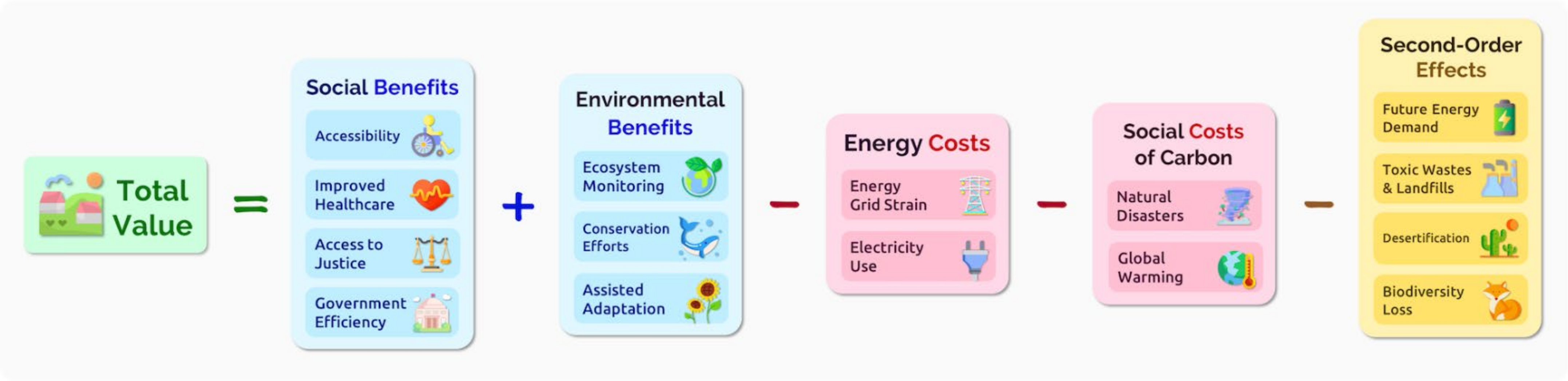
**Results show responses from teachers, principals, and district leaders.*

SOURCE: EdWeek Research Center survey, February 2022

2. Bias - Solutions

- Better community (education)
  - Commitment to moving beyond gender binary processing
  - More nuanced treatment of race
  - …
- Personalization the Study
- Involvement of experts from different backgrounds in the research process
  - Minorities, subcultures, people with disabilities, researchers from different countries, etc.
- Better data
- Better models
- Better norms allow AI owners to block specific content according to norms effectively.

3. Environment



Social Benefits

Accessibility | Improved Education | Away from harm | School Efficiency

[6]

Unfair environment

- The economic benefits and social costs of carbon may be unevenly distributed across communities.
    - The health effects of establishing a toxic waste site in Silicon Valley are unevenly distributed to socially disadvantaged groups
    - Taiwanese manufacturing pollution is associated with chronic health problems.

Environmental impact

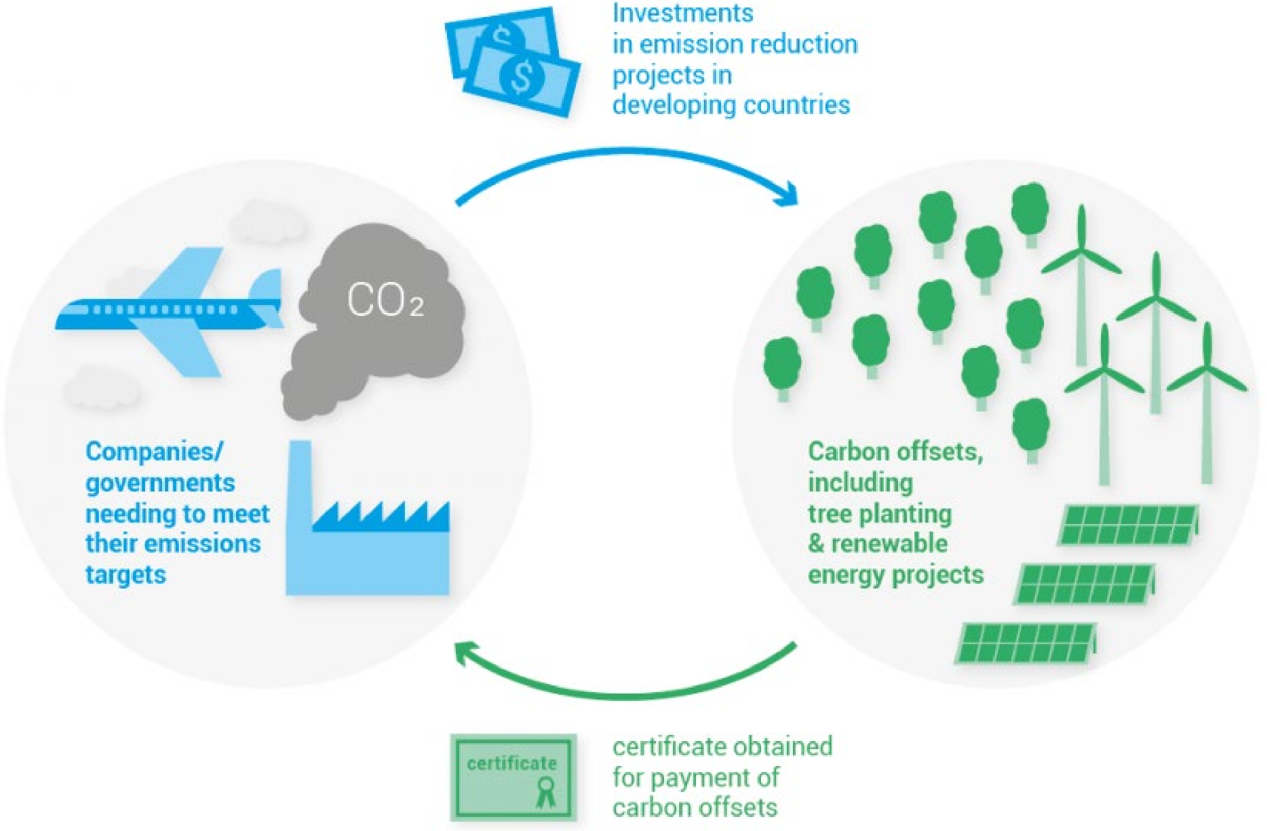
- Accelerating Deserts
- Rapid ecosystem change puts many species at risk of extinction
- Increased carbon emissions due to melting permafrost

Environmental impact

- Accelerating Deserts
- Rapid ecosystem change puts many species at risk of extinction
- Increased carbon emissions due to melting permafrost

Fraudulent carbon offsets

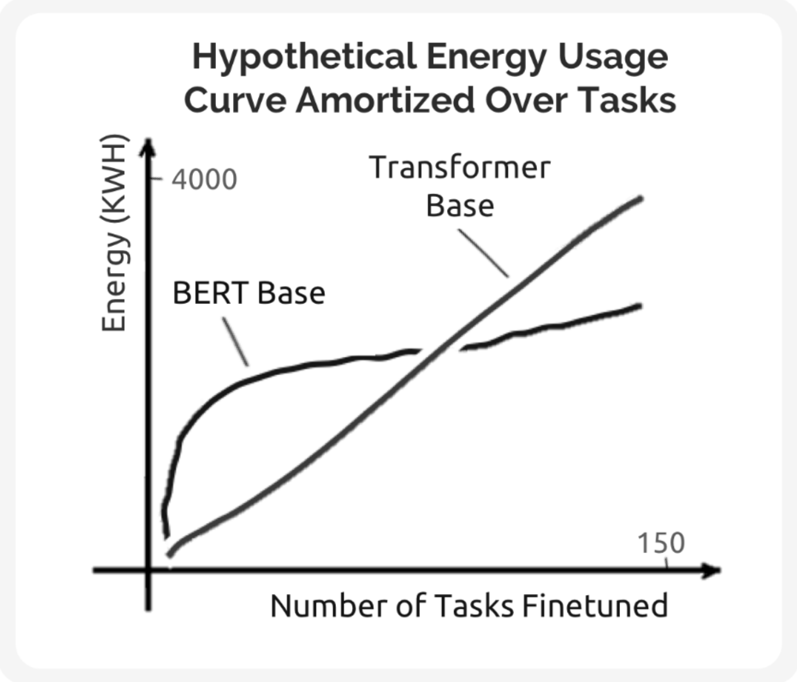
- Most carbon offsets are a **worse solution** than not emitting CO2, to begin with. Using carbon offsets may result in more carbon emissions than if the carbon had never been emitted in the first place.


Investments in emission reduction projects in developing countries
$
CO₂
Companies/ governments needing to meet their emissions targets
Carbon offsets, including tree planting & renewable energy projects
certificate
certificate obtained for payment of carbon offsets

https://www.unep.org/news-and-stories/story/carbon-offsets-are-not-our-get-out-jail-free-card

3. Environment - Solutions

- Activities in low carbon intensity areas
- Better Model architectures
  - Some models are more suitable for large scale and not power hungry.
- Distillation models
  - In [12], it was shown that we could make a small model perform as well as a model 2000 times larger by distilling it step by step.



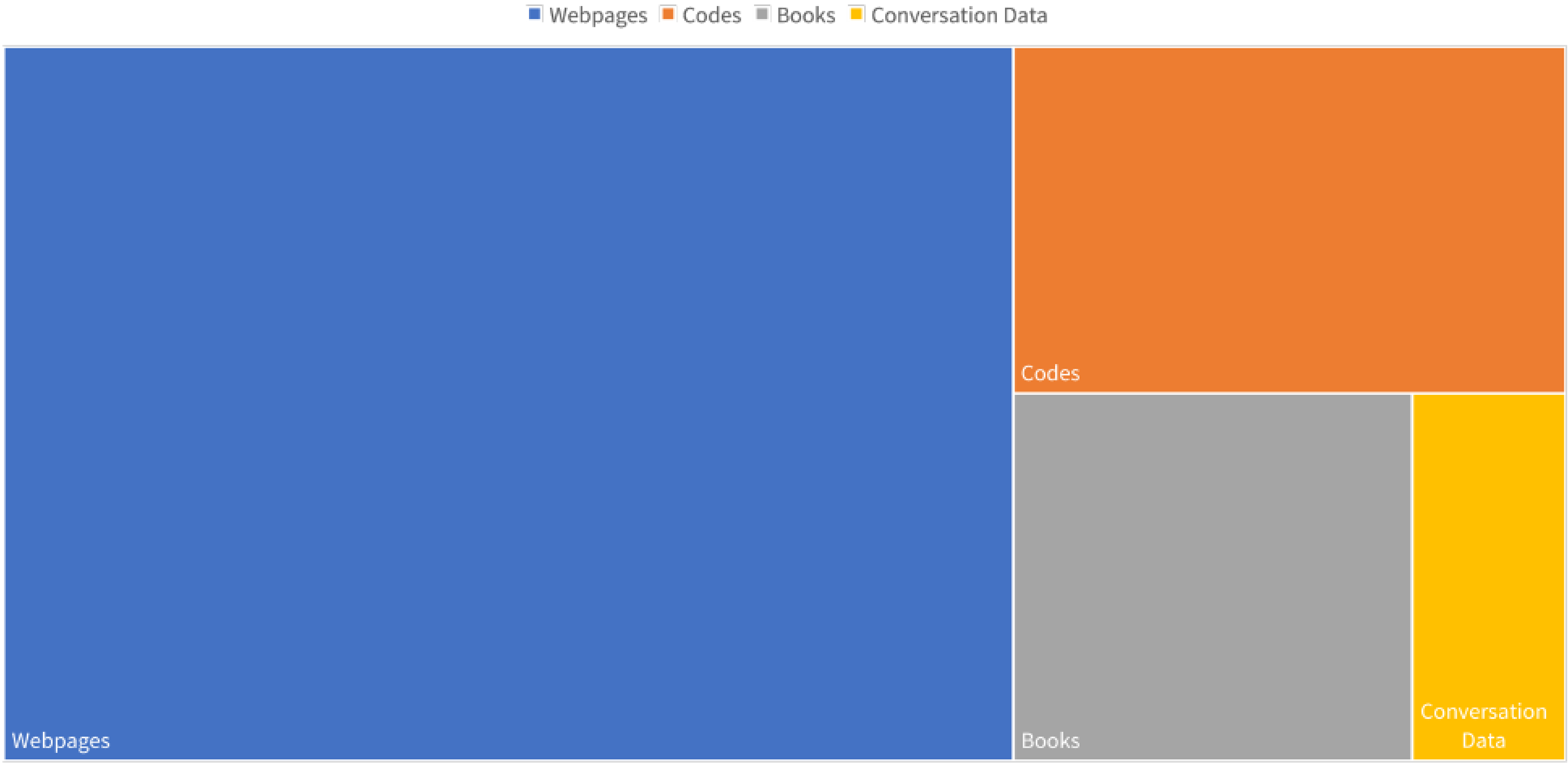
Why is it hard to be Ethical

- Data are biased
- Oversight of the education system is inadequate
- Hard to convince everyone
    - It is difficult for different education systems to agree on rule standards

Data Limitations, Bias, Representativeness

- Example: GPT-3
- The GPT-3 was trained using webpages, Reddit posts, and Wikipedia up to 2021.
  - The training data over-records statements from young users in developed countries.
  - More than half of the users in the Reddit forums are young men[1].
  - Only 8.8%-15% of Wikipedia users are female[2].

Reference: [1] Reddit news users more likely to be male, young and digital in their news preferences. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/
[2] Gender bias on Wikipedia. https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia



Ratios of various data sources in the pre-training data for

Adapted from W.Zhao et al., A Survey of Large Language Models, 2023

Noble goals are no substitute for ethical oversight

- As noted previously, there is no doubt that all respondents do wish to explore, better understand and engage with the ethics of the design and application of AI in educational contexts. However, some respondents appear to believe that we, as a community, are already ‘doing ethics’ by virtue of operating with best intentions in the educational domain, which is in and of itself ethical (e.g., “*Many of us got into AIED because we wanted to improve education* [...]. *Pursuing that goal is doing ethics* [...]. Yet, definitely, we could do more.” Koedinger). However, the reality is that “*no ethical oversight is required to deploy an eLearning system or an AIED system as part of the normal teaching process*” (du Boulay).

Conclusion

- Proactive intervention.
  - Preventative ethical risk reduction, starting with how models are developed and deployed.
  - **Ethical guidelines and standards**
- Proactive response.
  - Responding to known errors promptly, actively pursuing problems, resolving them quickly and accurately, and making changes in the future.
  - **Ongoing monitoring and evaluation**
- Community values(education).
  - Expand the diversity of researchers and model users. Build better community values and behavioral norms for AI faculty.
- **Regulatory framework**
  - Governments need to update existing laws or create new ones to address the unique challenges AI-generated content poses. This may include regulations regarding intellectual property, plagiarism, data protection, and privacy

Reference

[1] The 2 Sigma Probem: The Search for Methods of Group Instrction as Effective as One-to-One Tutoring

[2] https://openai.com/customer-stories/khan-academy

[3] https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/

[4] https://hai.stanford.edu/news/ai-will-transform-teaching-and-learning-lets-get-it-right

[5] Ethics of AI in Education: Towards a Community-Wide Framework

[6] On the Opportunities and Risks of Foundation Models

[7] Poesia, G. and Goodman, N.D. (2023). Peano: Learning Formal Mathematical Reasoning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, [online] 381(2251), p.20220044. doi:https://doi.org/10.1098/rsta.2022.0044.

[8] Anthony, T., Tian, Z. and Barber, D. (2017). *Thinking Fast and Slow with Deep Learning and Tree Search*. [online] arXiv.org. Available at: https://arxiv.org/abs/1705.08439.

[9] https://www.wearedevelopers.com/magazine/generative-ai-image-bias

[10] https://www.sohu.com/a/18373609_105067

[11] Leonardo Nicoletti and Dina Bass HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE https://www.bloomberg.com/graphics/2023-generative-ai-bias/?utm_medium=deeplink&leadSource=uverify%20wall

[12] Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y. and Pfister, T. (2023). *Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2305.02301.

[13] https://www.scribd.com/document/421898931/The-Risk-of-Racial-Bias-in-Hate-Speech-Detection#from_embed

[14] Vidgen, B. and Derczynski, L. (2020). Directions in Abusive Language Training Data: Garbage In, Garbage Out. *arXiv:2004.01670 [cs]*. [online] Available at: https://arxiv.org/abs/2004.01670.

[15] https://lilianweng.github.io/posts/2021-03-21-lm-toxicity/

[16] Zhou, K., Ethayarajh, K. and Jurafsky, D. (2021). Frequency-based Distortions in Contextualized Word Embeddings. *arXiv:2104.08465 [cs]*. [online] Available at: https://arxiv.org/abs/2104.08465.