Written Testimony of
Emily M. Bender
Professor of Linguistics
University of Washington

Joint Subcommittee Hearing on:
Balancing Knowledge and Governance:
Foundations for Effective Risk Management of Artificial Intelligence

before the
United States the House Committee on
Science, Space and Technology
Subcommittee on
Oversight & Investigations and
Research & Technology Subcommittee

October 18, 2023

Good morning, Chairmen Obernolte and Collins and Ranking Members Foushee and Stevens. Good morning distinguished Members of the Committee. Thank you for the opportunity to speak with you today about the risks associated with the technologies being marketed as "artificial intelligence". My name is Emily M. Bender, and I am a Professor of Linguistics at the University of Washington. This technological sector is presently the site of enormous influxes of capital and enormous concentrations of power. In order to channel the innovation taking place towards the benefit of society, we need thoughtful regulation that shores up our rights. I am very heartened to see your committee engaging in these important discussions.

A key step towards clear discussions of these topics is clear terminology. I find that the phrase "artificial intelligence" is best understood as a marketing term, and one which only muddies the waters. It is clearer to talk about automation. That helps us ask important questions like:

- What is being automated?
- Who is automating it and why?
- Who benefits from that automation?
- How well does the automation work in its use case that we're considering?
- Who is being harmed and what recourse do they need?
- Who has accountability for the functioning of the automated system?
- What existing regulations already apply to the activities where the automation is being used?

Types of automation sometimes called "AI" include the following (among others):

- Automated **decision systems** and **decision support systems**, such as those used to help relevant officials decide how to set bail, which loans to approve, which resumes to look at more closely, and how to allocate social benefits.

- Automated **classification**, of images, text, or datafied representations of people. Applications of this include benign use cases like helping digital cameras keep the focus on people in a photo, but also oppressive uses of facial recognition technology for surveillance purposes and systems which purport to detect the sound of gunshots.

- Automated **recommender systems**. These are systems which automate the choice of what information to present to a person such as in algorithmic social media feeds.

- Systems providing an **application programming interface layer** for access to human labor. Examples here include services like Uber and Lyft but also platforms like Amazon Mechanical Turk. In all cases, the interface layer lets external actors treat the human labor almost as software components.

- Automated **translation** of information between formats. This category includes applications such as automatic transcription (speech to text), speech synthesis (text to speech), machine translation, code generation from natural language descriptions, and image style transfer.

- **Synthetic media machines**, like ChatGPT, DALL-E, etc. These are systems which create text, images, video or audio based on input prompts.

I'd like to say a few more words about ChatGPT, because it is important to understand how it produces the illusion of understanding (and therefore the illusion of intelligence). ChatGPT is fundamentally a language model, and more precisely a model of the text it is trained on. That means that its only task is to repeatedly produce a likely next word, given some input word sequence or "prompt", according to the distributions in its training text. Because it is trained with enormous amounts of text, it can produce plausible seeming output on nearly any topic. But that output comes without any commitment to its contents. At present, no one is actually accountable for the synthetic media being spilled into our information ecosystem.

At the same time, we can't help but make sense of language that we encounter. We do that by instinctually imagining the point of view of the person speaking or writing in order to infer what they wanted us to understand by saying what they said.[1] Prior to the advent of ChatGPT, if we encountered some coherent-seeming text, it was safe to assume that it came from some person or group of people who were using it to communicate something. So it is only natural that when we see the output of ChatGPT, we imagine that the computer behind it is doing the same thing — when it emphatically is not.

In short, we relate to ChatGPT analogously to how one plays with the old Magic 8 Ball toy: That toy provides answers, most of which only make sense for yes-no questions. If you ask "What should I have for lunch?" and it replies "Signs point to yes", the conversation is incoherent. Accordingly, in playing with the toy, we quickly learn to use only yes-no questions so that we can

[1] Clark, H.H. (1996). *Using Language*. Cambridge University Press.

make sense of the output. With ChatGPT, we are doing a similar exercise. For ChatGPT, the input is simply a string of words that form the prompt for it to continue, with likely next words. For us as users, we frame our input to ChatGPT so that we can make sense of the output as if it were an answer.

Another part of the illusion that ChatGPT "understands" comes from the way it can be used for translation tasks, plus the fact that its training data includes natural language text paired with related computer code.[2] A translation task maps one representation to another, such as English text to French text, or English text to computer code. When people do translation tasks, we centrally use our understanding of the content of what we are mapping. And when we write computer code, it often isn't a translation task for us at all: rather, we reason about a problem and then break it down into smaller pieces. So when we see a computer translating English text to computer code, this reinforces an illusion of "understanding" and "intelligence".

If we don't account for our own sensemaking abilities when we try to evaluate these technologies, we can end up concerned with fantasy scenarios instead of real-world risks and harms. Sometimes in conversations about the risks of so-called "AI", you'll hear a strong focus on something called "existential risk" or the idea that the "AI" might become sentient and turn on us.

It is important to know that discussions of fantastical, malevolent, autonomous thinking machines are a deflection and a distraction. They are a deflection because they locate the potential for harm in the technology itself, rather than in the choices of human actors. They are a distraction because they point your attention towards imaginary scenarios and away from the actually occurring harms and real-world risks.

Those real risks and harms include things like the following:

- The pollution of the information ecosystem with synthetic media[3]
- The reproduction of biased and discriminatory patterns in advertizing,[4] search results,[5] generated media, and decision making, including in high-stakes scenarios in housing, employment and the justice system[6]
- The exploitation of workers and data theft underlying the creation of these systems[7]
- Substantial environmental impact, both in carbon footprint and water usage[8]

[2]OpenAI has been anything but transparent about the training data it uses, but the presence of such training data can be inferred from system output.

[3]Shah, C. and Bender E.M. (Under review). Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? https://bit.ly/Env_IAS

[4]Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM* 56.5:44-54.

[5]Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press

[6]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[7]https://time.com/6247678/openai-chatgpt-kenya-workers/
https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483

[8]https://www.theguardian.com/technology/2023/aug/01/techscape-environment-cost-ai-artificial-intelligence

- The oppressive surveillance of people:[9] as citizens, as migrants, as workers, as the formerly and currently incarcerated

- The fraying of social services: when automated allocation systems being used to reduce the amount of support offered,[10] or via the replacement quality health care, education and other social benefits with synthetic text and other automated systems

- And finally, the use of automation to replace skilled jobs with deskilled, precarious gigwork[11]

These harms are real and present, but we are not powerless. I believe that an effective regulatory remedy would include the following elements:

- Requirements of transparency about the fact of automation, including of synthetic media. We should always know when we are encountering the output of a synthetic media machine and we should always know when automated pattern matching has been involved in decisions about us.

- Requirements of transparency about the data systems are trained on and their resulting performance. There are many detailed proposals for how to thoroughly document both datasets used in training these systems and the resulting models.[12]

- Clear accountability for system outputs. Computer programs are not the kind of thing that can take responsibility.[13] When automation is involved in producing some kind of output (decision, recommendation, synthetic media) it should always be clear which person, corporation or other organization is accountable for any harm caused by that output.

- Recourse for people who are adversely affected by automated systems. Automated systems have the ability to scale harm and are also inflexible — whereas the humane and fair treatment of people demands discretion.[14]

[9] Ajunwa, I. (2023). *The Quantified Worker: Law and Technology in the Modern Workplace.* Cambridge University Press.
Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

[10] Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. MacMillan Publishers.

[11] Gray, M. and Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt.

[12] Bender, E.M., Freidman, B., and McMillan-Major, A. (2021). A guide for writing data statements for natural language processing. `https://techpolicylab.uw.edu/data-statements/`.
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12):8692.
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, p.220–229.

[13] Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W.H. Freeman.

[14] Alkhatib, A. (2021). To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, p.1–9.

- National funding for humanities and social science research into the impact of technology on society. The expertise required to understand how automation can adversely impact people and society lies in the fields of scholarship which engage with social structures and the human condition. To effectively understand the risks of so-called "artificial intelligence" we need vibrant and well-supported scholarly communities in these fields.

Thank you again for your attention and for the opportunity to speak with you today about these important matters.