BLOG@CACM

The *Communications* website, https://cacm.acm.org, features more than a dozen bloggers in the BLOG@CACM community. In each issue of *Communications*, we'll publish selected posts or excerpts.

twitter
Follow us on Twitter at http://twitter.com/blogCACM

DOI:10.1145/3613250 https://cacm.acm.org/blogs/blog-cacm

Controlling AI

Gary Marcus considers it appropriate that governments are stepping up on artificial intelligence, but finds some of the resulting signals deeply worrisome.

Gary Marcus

Two Models of AI Oversight — and How Things Could Go Deeply Wrong

https://bit.ly/3Qnxd9A

June 12, 2023

Originally published on The Road to AI We Can Trust (http://bit.ly/3juuD3j)

The Senate hearing that I participated in a few weeks ago (https://bit.ly/44QxHt1) was, in many ways, the highlight of my career. I was thrilled by what I saw of the Senate that day: genuine interest and genuine humility. Senators acknowledged that they were too slow to figure out what do about social media, that the moves were made then, and that there was now a sense of urgency. I am profoundly grateful to Senator Blumenthal's office for allowing me to participate and tremendously heartened that there was far more bipartisan consensus around regulation than I had anticipated. Things have moved in a positive direction since then.

But we haven't landed the plane yet.

Just a few weeks earlier, I had been writing in this Substack and in *The Economist* (https://econ.st/3Kpzm0C, with Anka Reuel) about the need for an international agency for AI. To my great surprise, OpenAI CEO Sam Altman told me before the proceedings began that he was supportive of the idea. Taken off guard, I shot back, "Terrific, you should tell the Senate," never expecting that he would. To my amazement, he did, interjecting, after I raised the notion of global AI, that he "wanted to echo support for what Mr. Marcus said."

Things have in many ways moved quickly since then, far faster than I might have ever dreamed. In 2017, I proposed a CERN for AI in *The New York Times* to relatively little response. This time, things (at least nominally) are moving at breakneck speed. Earlier this week, British Prime Minister Rishi Sunak explicitly called for a CERN for AI, as well something like an IAEA for AI, all very much in line with what I and others have hoped for. Earlier ... President Biden and Prime Minster Sunak agreed ... publicly, "to work together on A.I. safety."

All that is incredibly gratifying. And yet ... I am still worried. Really, really worried.

What I am worried about is regulatory capture (https://bit.ly/3KpD7mF); governments making rules that entrench the incumbents, whilst doing too little for humanity.

The realistic possibility of this scenario was captured viscerally in a sharp tweet from British technology expert Rachel Coldicutt:

I had similar pit-of-my-stomach feeling in May after VP Kamala Harris met with some tech executives, with scientists scarcely mentioned.

Putting it bluntly: if we have the *right* regulation, things could go well. If we have the wrong regulation, things could badly. If Big Tech writes the rules, without outside input, we are unlikely to wind up with the right rules.

In a talk I gave to the IMF, I painted two scenarios, one positive, one negative:

The Positive Future

- **2023: Global AI agency was formed, and AI was thoughtfully regulated.**
- **2024: Responsible AI became a prestigious profession.**
- **2025: New companies and new tech emerged.**
- **2026: AI become more efficient, in terms of both data and energy.**
- **2026–2029: AI begin to contribute massively to the world, addressing climate change, medicine, eldercare, and many more.**

The Bleak Future

- **2023: Conflicts over which risks to address precluded anything from happening: "AI Safety" and "AI Ethics" people couldn't agree on anything, either on terms of problems or solutions; Congress gave up in disgust.**
- **2023: We got stuck on LLMs and never invented a better, more reliable, more efficient tech.**
- **2025: A small number of companies quickly become far more powerful than states, running the world as they please, shutting out all competition with ill-conceived regulation of their own devising.**
- **2025: Cybercrime syndications and big companies begin an epic battle, reminiscent of drug cartel wars.**
- **2027: Increasingly powerful AI systems are constructed and quickly become weaponized; large numbers of people are killed in deadly conflicts, both deliberate and accidental.**
- **2029: Employment crashes, widespread unrest. Multiple civil wars. Anarchy.**

We still have agency here; we can still, I think, build a very positive AI future.

Yet much depends on how much the government stands up to Big Tech, and a lot of that depends on having independent voices—scientists, ethicists, and representatives of civil society—at the table. Press releases and photo opportunities that highlight government officials hanging out with the tech moguls they seek to regulate, without independent voices in the room, send entirely the wrong message.

The rubber meets the road in implementation. We have, for example, Microsoft declaring right now that transparency and safety are key. But their current, actual products are definitely not transparent, and at least in some ways, are demonstrably not safe.

Bing relies on GPT-4, and we (that is, in the scientific community) do not have access to how GPT-4 works, and we do not have access to what data it is trained on (vital, since we know that systems can bias, for example, political thought and hiring decisions based on those undisclosed data)—that is about as far away from transparency as we could be.

We also know, for example, that Bing has defamed people, and it has misread articles as saying the opposite of what they actually say, in service of doing so. Recommending (*New York Times* technology columnist) Kevin Roose get a divorce was not exactly competent, either. Meanwhile, ChatGPT plug-ins (produced by OpenAI, which they have a close tie with) open a wide range of security problems: They can access the Internet, read and write files, and impersonate people (for example, to phish for credentials), all alarms to any security professional. I don't see any reason to think these plug-ins are, in fact, safe. (They are far less sandboxed and less rigorously controlled than Apple app store apps.)

This is where the government needs to step up and say, "Transparency and safety are indeed requirements; you've flouted them; we won't let you do that anymore."

We don't need more photo opportunities, we need regulation—with teeth.

More broadly, at an absolute minimum, governments need to establish an approval process for any AI that is deployed at large scale, showing that the benefits outweigh the risks, and to *mandate* post-release auditing—by independent outsiders—of any large-scale deployments. Governments should demand that systems only use copyrighted content from content providers that opt in, and that all machine-generated content be labeled as such. And governments need to make sure there are strong liability laws in place to ensure that if big tech companies cause harm with their products, they be held responsible.

Letting the companies set the rules on their own is unlikely to get us to any of these places.

In the aftermath of the Senate hearings, a popular sport is to ask, "Is Sam Altman sincere, when he has asked for government regulation of AI?"

A lot of people doubted him; having sat three feet away from him, throughout the testimony, and watched his body language, I actually think that he is at least in part sincere, that it is not just a ploy to keep the incumbents in and small competitors out, that he is genuinely worried about the risks (ranging from misinformation to serious physical harm to humanity). I said as much to the Senate, for what it's worth.

But it *doesn't matter whether Sam is sincere or not*. He is not the only actor in this play; Microsoft, for example, has access, as I understand it, according to rumor, to all of OpenAI's models, and can do as they please with them. If Sam is worried, but Nadella wants to race forward, Nadella has that right. Nadella has said he wants to make Google dance, and he has.

What really matters is what governments around the world come up with by way of regulation.

We would never leave the pharmaceutical industry to entirely self-regulate itself, and we shouldn't leave AI to do so, either. It doesn't matter what Microsoft or OpenAI or Google say. It matters what the government says.

Either they stand up to big tech, or they don't; the fate of humanity may very well rest in the balance.

Gary Marcus (@garymarcus), scientist, bestselling author, and entrepreneur, is deeply concerned about current AI but really hoping we might do better. He spoke to the U.S. Senate on May 16, 2023, and is the co-author of the award-winning book *Rebooting AI*, as well as host of the new podcast *Humans versus Machines*.