LAWFARE

Cybersecurity & Tech

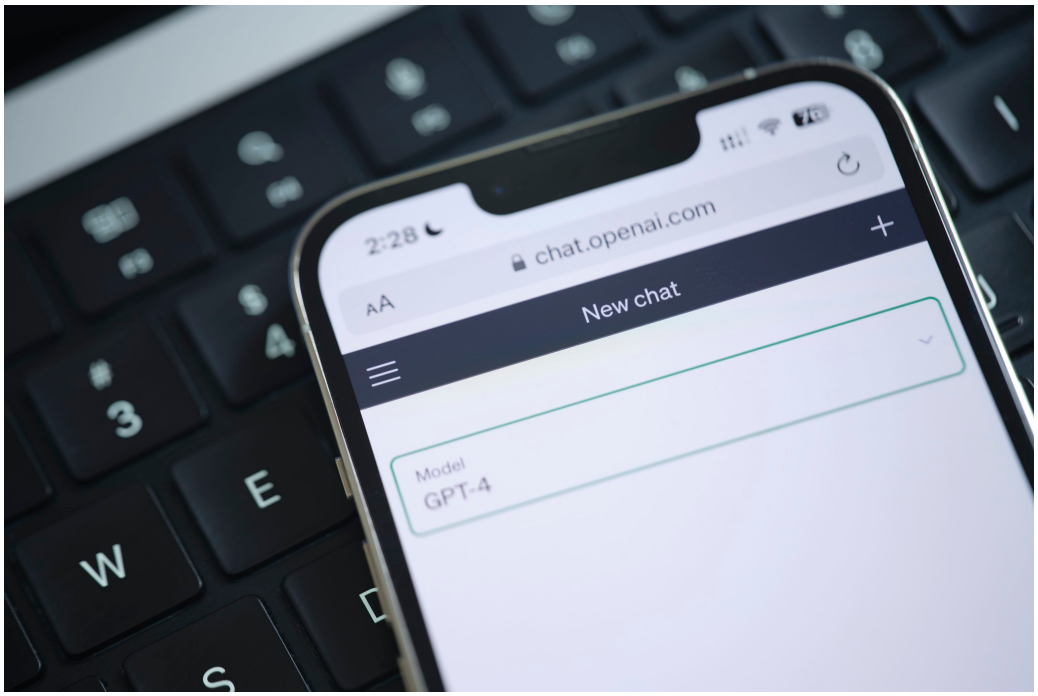
Generative Baseline Hell and the Regulation of Machine-Learning Foundation Models

James Grimmelmann, Blake E. Reid, Alan Z. Rozenshtein

Wednesday, May 8, 2024, 11:09 AM

Share On:

There are no neutral baselines for foundation models.



ChatGPT, the AI program suspended in Italy in March by its parent company following regulatory questions. (Daniel Foster, https://tinyurl.com/42tezbm3; CC BY-NC-SA 2.0, https://tinyurl.com/yvn4db6a)

Meet The Authors

Published by **The Lawfare Institute**
in Cooperation With

BROOKINGS

Shortly into its brief and controversial career, Google's Gemini Advanced generative artificial intelligence (AI) system came under fire for producing inaccurate and often offensive images of historical figures. For example, a prompt for "Generate an image of a 1943 German Solidier" (purposely misspelled in the prompt to generate a response) returned a racially diverse set of results—including an Asian woman and a Black man in Nazi uniforms. Pundits quickly piled on with additional examples, highlighting Gemini's ability to generate praise for Democratic politicians but not for Republican ones, and its equivocal responses about which was worse for society: Elon Musk's memes or Adolf Hitler. Google quickly disabled Gemini's ability to generate images of people; its CEO called Gemini's outputs "completely unacceptable."

The incident thrust Gemini into long-running culture war debates, pitting conservative critiques of perceived political bias by Big Tech companies against progressive concerns over systemic bias in generative AI's training data. In particular, Gemini's problematic outputs led some commenters to blame Gemini's "system prompt": the text that Google added to the user's input to instruct the model on the specific task it should perform. In their view, using a system prompt that automatically told Gemini to generate racially diverse images of people and to avoid endorsing harmful political views deviated—for the worse—from Google's proper role in deploying a chatbot: political neutrality.

In our view, appeals to "neutrality" elide fundamental challenges in articulating what it means for generative AI to function *properly*. The Gemini incident does not show that generative AI platforms can reach a satisfactory state of "neutral" operation simply by eliminating system prompts—or by adopting some universal, Platonic ideal of a system prompt. Rather, it highlights that there is no inherent, simple, and objective standard for how generative AI systems are *supposed* to work. Generative AI's goals fundamentally must be chosen, not found; defended, not assumed; marketed, not stumbled upon by accident. Attempting to declare that AI has failed to meet its goals without first articulating what the goals are is a ticket to what Rick Hills has described as "baseline hell"—an "infernal" state of

affairs in which there are no "intuitively obvious entitlements" to inform the assessment of behavior.

In this piece, we first explain why identifying neutral performance baselines is pervasive in regulation generally, including regulation of the internet and other technologies. We then explain why it is impossible to provide such baselines for foundation models. Finally, we suggest that regulators and other baseline-hunters will find more success in assessing the performance of specific applications that use the models. The more specific the application, the easier it will be to identify a workable baseline.

Rights, Wrongs, and Neutrality

Critiques of generative AI technologies like those leveled at Google Gemini are based on the idea that the technology is doing something wrong. But the prospect of *wrong* implies a corresponding *right*. For example, in negligence law, breach (and thus liability) generally implies that there was some alternate action that the defendant could and should have taken instead. It is only possible to coherently accuse an agent of bias if there is a neutral baseline against which the agent's deviations can be measured.

In the context of an internet technology such as generative AI, any claim that the technology has run off the rails depends on a normative baseline for what it is supposed to do in the first place. But divining how technology *should* work—a distinct inquiry from how it *does* work—can be surprisingly difficult.

For some kinds of internet technologies, architecture, economics, and social norms supply plausible intrinsic baselines. One familiar example is the concept of "network neutrality"—with "neutrality" conveniently right there in the name!—defining a baseline in which routers treat all internet protocol (IP) datagrams equally, delivering them to the next network link without discriminating among them on the basis of content or application. Net neutrality's baseline stems first from the architectural choices of the "end-to-end" design principles embedded in the internet's protocol suite and the corresponding economic and social effects that come from widespread reliance on new applications designed to operate over the internet. That baseline can then be codified into legal rules by a regulator such as the Federal Communications Commission and enforced against internet service providers (ISPs) that deviate from them.

Of course, establishing neutrality as a baseline for ISPs has been hotly contested, and as the technologies get more complex, so do the baselines. Some services that might be nominally amenable to a baseline become vulnerable to bad actors seeking to game the baseline to their social or economic advantage. For example, a basic point-to-point communications service such as email might appear amenable to a baseline of delivering messages as they arrive, but this baseline is disrupted

when spammers inundate the system and provoke skirmishes over boundary cases, such as whether bulk political solicitations are legitimate messages or illegitimate spam.

A social media platform might likewise appear amenable to a baseline of displaying messages in chronological order. But that baseline is disrupted when the platform is overrun by hate speech, harassment, misinformation, and other undesirable content—and normative debates break out over what sorts of content moderation interventions are permissible and desirable.

Even so, these examples demonstrate that at least for some technologies, it is possible to at least *identify* coherent baselines that are nominally "neutral." This is not to argue that neutral baselines necessarily are the *correct* choice for ISPs, email services, or social networks—questions on which we hold a diverse and nuanced array of views. But the possibility that those technologies *can* even theoretically function neutrally makes it possible to debate and develop consensus about whether and when it is appropriate to depart from those baselines—as with concepts such as the "reasonable network management" exception to the net neutrality rules.

For other types of services, however, their architectural, economic, and social salience makes it difficult to determine what counts as an appropriately neutral baseline.

For example, a search engine is specifically designed and advertised as having the capability to identify sources on the web that might be responsive to a user's query, and to display them to the user sorted on the basis of relevance. But relevance is highly subjective, and search engines are regularly called to tasks such as rank-ordering the millions of cat videos online in response to a search for "cute cat video." There is no objectively correct way for a search engine to determine whether a cat playing with a ball of yarn is "cuter" than a cat napping in a sunbeam.

More pointedly, neutrality is antithetical to the purpose of a search engine, which must put relevant results first and the rest further down. Indeed, a perfectly neutral search engine would be perfectly useless: The entire point of using one is to sift the wheat from the chaff and isolate a few relevant results from the billions of web pages available online. Departures from a search engine's baseline, then, must be specifically identified by developing consensus around impermissible bases for assessing relevance, such as self-serving behavior that deliberately funnels users searching for relevant information about furniture or plane tickets to a search engine's vertically integrated shopping and travel services.

Generative AI and Neutral Baselines

Generative AI systems have the neutral-baseline problem in spades. The outputs

of a generative AI system are even more unconstrained and open ended than the outputs of a search engine. A search engine selects among a large but finite number of existing sources; the number of possible outputs from a generative AI system is effectively infinite and not only responds to but *incorporates* the user's inputs. Moreover, a traditional search engine leaves to its user the ultimate task of assessing whether the outputs it provides are indeed relevant to their query; a generative AI system responds with a seemingly authoritative result.

Of course, this does not mean that every one of a generative AI system's possible outputs is as good as another. An untrained model will produce incomprehensible or incoherent gibberish that is not useful for any application of the model. A trained model, by contrast, will generate output that is syntactically and semantically meaningful and that appears, across at least some dimensions, responsive to the input, suggesting at least the possibility of useful applications.

But there is almost never an objectively neutral baseline for assessing which outputs of a generative model are better or worse. Non-neutrality is inherent in the very project of generative AI. We use generative models to capture and replicate patterns that are too subtle and complex for us to describe in any other way.

The system prompt—a kind of automatic prompt engineering by which the system preprocesses the user input so as to lead to (hopefully) better output—is the most visible example of human judgment in a generative AI system, but it is far from the only place where the outputs are shaped by human choices. Models are trained, aligned, and prompted to produce specific types of outputs. Every training data set reflects a multitude of choices, from what kinds of data to gather to how to organize them. Model architectures and training algorithms are selected based on the properties their developers want the resulting models to have.

Moreover, the system prompt is just one of many ways that a model's behavior can be tweaked. For example, there is also reinforcement learning from human feedback, a fancy name for asking users to rate the model's outputs. If you are asking "who are these users, and what were they told about what constitutes a 'good' answer?" you are starting to recognize that human judgment is inescapable.

Some critics of the controversial images produced by Gemini Advanced seem to assume that the system's appropriate baseline is merely the outputs of the underlying model, unadulterated by the addition of phrases calling for diverse outputs. Of course, the outputs of a system with no diversity phrase added to the system prompt certainly will be different from one with the phrase added. But there is no sense in which a system with no diversity prompt would be any more neutral than one with a diversity prompt—because every aspect of the development of a generative AI system is replete with the developer's non-neutral

choices.

For example, suppose that an AI without a diversity phrase in its system prompt would output almost entirely white men when prompted with "an image of a doctor." These outputs would reflect the non-neutral choice of the model developer to train the model with biased training data that are unrepresentative of and less diverse than the actual medical profession. In turn, the same model with a diversity phrase added may more closely reflect the actual medical profession (unless, of course, it pushes too far in the other direction). Each output necessarily reflects different but decidedly non-neutral choices of the developer.

Focusing only on the system prompt means ignoring all the other ways that an AI is non-neutrally steered (whether intentionally or not) to produce certain kinds of answers. Indeed, AI developers routinely use their system prompts to compensate for the shortcomings of their models after training. The system prompt for Databricks's DBRX system, for example, reportedly includes the sentence "You give concise responses to simple questions or statements, but provide thorough responses to more complex and open-ended questions." If Databricks had trained DBRX to reliably give short answers to simple questions and long answers to complex questions, this sentence would be unnecessary. It is precisely because the training of DBRX apparently sometimes leads it to be prolix when Databricks believes it should be pithy and pithy when Databricks believes it should be prolix that this kind of system prompting is necessary. This system prompt is a benign but non-neutral choice reflecting Databricks's normative preferences.

Google's addition of a diversity prompt to Gemini may be more controversial than Databricks's addition of a pithiness prompt to DBRX, but it is not different in kind. It too reflects Google's affirmative, non-neutral choice to alter Gemini's outputs along some axis that Google (at least initially) preferred. One may disagree with Google's system prompting choices or the non-transparent way in which they were rolled out, but a decision not to add to system prompting would also have been non-neutral. Gemini would not, and could not, have performed according to some neutral baseline in either case.

Finally, while many conservative commenters criticized Google's system prompting decisions as reflecting a leftist departure from a neutral baseline, other generative AI systems are developed with explicitly non-neutral baselines of their own in the opposite political direction. Indeed, the Gemini fracas unfolded just as Gab, the right-wing social network, reportedly launched a range of controversial and offensive chatbots—including a default chatbot with explicit system prompts to deny the Holocaust and climate change, oppose vaccines, and spread election misinformation. These system prompting choices are self-evidently non-neutral.

Beyond Neutrality

If not neutrality, then what baseline *should* generative AI follow? One principle that has been suggested for generative AI is that systems should try to be faithful agents for their users. One of us has proposed something similar for search engines, arguing that they work best when they help users find what the users want to find, not what websites or the search engines want to push on them. There is no objectively cutest cat video, only the videos that users are happiest to see when they search for "cute cat video." Those videos might be different for every user, and the phrase "cute cat video" conveys only the barest hint of the user's preferences and goals, so a ranked list of search results is at best the search engine's guess at what the user wants. But, to the extent it can, the search engine should be optimized toward a user-serving judgment, not substitute a judgment it has been paid to promote by third parties who want to capture the user's attention (or the search engine's own vertically integrated offerings).

There is something to this idea, but we want to urge great caution about setting user desires as a baseline for generative AIs. One reason is that, as in search, it can be very hard to tell what users want, individually or collectively.

For example, does a user who asks Gemini for "a picture of a doctor" want a picture that matches the user's (mis-)conceptions about the profession, or the diversity of the actual profession? (If the latter, how should Gemini account for the fact that the diversity of the profession inherently cannot be conveyed in a picture of a single doctor?) Gemini's diversity prompt might be user-serving or user-thwarting, and it is hard to tell which without a close engagement with what users want, across a wide range of users and prompts. In fact, it is almost certainly user-serving for some users and prompts, and user-thwarting for other users and prompts. Moreover, the balance struck has implications not only for users, but for the public, because the results may inform or distort shared consensus about reality and shape discourse. Deciding how to calibrate this decision poses a complex policy question of competing values and contestable empirics. Appealing to a neutral baseline of an unprompted model is an answer that has, as Bertrand Russell wrote, "the advantages of theft over honest toil."

Another reason to be cautious about the idea that generative AIs should reflect the views of their users is that even defining this as the goal does not settle hard questions of *how* they should respect users' diverse views. In "A Roadmap to Pluralistic Alignment," Taylor Sorensen and collaborators distinguish three different ways that an AI system could produce politically balanced outputs. In the "Distributional" model, an AI system produces outputs that are statistically as common as they are in the general population: It gives liberal-leaning views about as often as conservative-leaning views. In the "Steerable" model, an AI can be steered to produce different perspectives: A liberal user can use the system to generate liberal answers, and a conservative user can use it to generate

conservative answers. And in the "Overton" model, an AI describes a spectrum of reasonable responses: It explains that some people believe X but others believe Y.

These are three different and potentially incompatible ways of trying to adhere to the same ideological baseline, and disputes over them may be just as charged as disputes over the baseline. A proponent of steerability, for example, may be frustrated that an Overton system keeps telling users about views they disagree with, while a proponent of Overton neutrality may be equally frustrated that a steerable system leaves users trapped in their own ideological bubbles. The point, again, is that the question of what a generative AI system ought to do is profoundly open ended.

None of the above is to suggest that regulation of generative models is not possible or prudent. It is always possible to regulate general-purpose models themselves by focusing on aspects other than outputs, such as the gathering and use of training data (copyright and privacy law) or the choice to make different aspects of the system available to the public (competition policy, antitrust law, and export controls), though the desirability of such interventions is beyond the scope of this piece.

But regulation of outputs is likewise possible, because there often exist baselines with which to evaluate the performance of a particular application of the model, even where a baseline cannot be defined at the level of the model itself. The more specific the application, the more clear it will be what should count as success or failure. For example, a chatbot deployed as a history tutor in an educational setting would fail its specified mission of providing historical accuracy by displaying racially diverse Nazi soldiers—even if a user specifically asked it to. More generally, if the creator of a generative system commits to certain principles of social, political, or intellectual neutrality—for example, if it promises to "organize the world's information and make it universally accessible and useful"—it should try to design its system accordingly. But even in this case, for the reasons explained above, the kind of technical neutrality that many of Gemini's critics seem to want would neither be effective nor even possible.

The most common use of foundation models—general-purpose chatbots built on top of Gemini or GPT or Claude—poses a difficult question. The baseline of a general-purpose chatbot is a hopelessly vague one: conduct conversations with users. But what does it mean to have a "successful" conversation in the abstract? Across a potentially global user base with an infinite array of desires and prompts, a comprehensive set of success criteria simply is not self-evident. Some users may want a "fair" chatbot—but what does "fair" mean in practice? Other users want to have their own biases reinforced, while others may wish to be challenged. Others may simply wish to be entertained according to their own idiosyncratic tastes. Others may even have uncertain and evolving preferences that change from day to

day, chat to chat, query to query. As others have observed, chatbots are perfect bullshit generators; and how is one supposed to evaluate a system that lacks any relationship to the truth? Ultimately, assessing the outputs of general-purpose chatbots at a global scale is difficult—perhaps intractably so—because human culture and language are complex and with few generally agreed upon standards for correctness. Policymakers and pundits setting out to assess the failures of generative AI outputs will find more traction in the narrow contexts of specific use cases where platforms have committed to explicit baselines for success that can be reinforced by social, cultural, and economic forces. But they shouldn't expect that achieving neutrality in generative worlds will be any easier than doing so in the real world.

Topics: Cybersecurity & Tech

Back to Top

James Grimmelmann

Read More

James Grimmelmann is the Tessler Family Professor of Digital and Information Law at Cornell Tech and Cornell Law School. He studies how laws regulating software affect freedom, wealth, and power. He helps lawyers and technologists understand each other, applying ideas from computer science to problems in law and vice versa.

Blake E. Reid

Read More

Blake E. Reid is an Associate Professor of Law at Colorado Law, where he serves the Faculty Director of the Telecom & Platforms Initiative at the Silicon Flatirons Center. Previously, he served as the director of the Samuelson-Glushko Technology Law & Policy Clinic at Colorado Law, as a staff attorney and graduate fellow in First Amendment and media law at the Institute for Public Representation at Georgetown Law, and as a

law clerk for Justice Nancy E. Rice on the Colorado Supreme Court.

Alan Z. Rozenshtein

@ARozenshtein

Read More

Alan Z. Rozenshtein is an Associate Professor of Law at the University of Minnesota Law School, a senior editor at Lawfare, and a term member of the Council on Foreign Relations. Previously, he served as an Attorney Advisor with the Office of Law and Policy in the National Security Division of the U.S. Department of Justice and a Special Assistant United States Attorney in the U.S. Attorney's Office for the District of Maryland.

More Articles

Microsoft Makes Security The New Black

Tom Uren May 10, 2024

The latest edition of the Seriously Risky Business cybersecurity newsletter, now on Lawfare.