Comments on the NTIA AI Open Model Weights RFC

ROBERT BRENNAN

MAR 8, 2024

Share

In response to an Executive Order from President Biden, The U.S. Department of Commerce (DoC) has asked the public for comments on "Open-Weight AI Models"—models like LLaMa, Stable Diffusion, and Mixtral—which are freely distributed to the public. They are considering blocking access these models in order to prevent abuse.

This would be a terrible mistake.

PSA: You can send the DoC your comments by clicking the "Comment" button on regulations.gov. Feel free to copy or paraphrase my comments below.

The deadline is March 27, 2024.

Background

The most sophisticated AI being developed today is closed-source (including, ironically, most of OpenAI's work). Scientists, engineers, and the public have no view into the inner workings of these algorithms. If we want to use them, we're forced to enter into a legal and/or financial agreement with the creator. Our activity is tightly restricted and monitored.

This is mostly normal. Corporations develop new technology, then control its use. ChatGPT's closed-source nature isn't meaningfully different from that of Google

Search or Snapchat.

But competitors are on the rise, and many of them are taking a more open approach. Meta, Stability AI, Mistral, and other companies are giving their work away for free, allowing us to run state-of-the-art models on our own hardware, outside the reach of anyone's oversight or control.

This, too, is mostly normal. People develop new technology, then give it away for free, since it doesn't cost anything to distribute. The open nature of Mixtral or LLaMa isn't meaningfully different from Linux or Signal.

The only difference with AI, is that AI is immensely powerful. The government is rightfully concerned about what will happen when advanced AI is widely available.

To put it bluntly: closed models allow for centralized control. Open systems, meanwhile, are free from oversight, and any controls can be trivially bypassed (see, e.g. How to Remove Stable Diffusion's Safety Filter in 5 Seconds).

So it's easy to see why the government might want to prevent the distribution of open models.

The Wrong Solution

The biggest threats posed by AI come not from individuals, but from corporations and state-level actors. And they will have unfettered access to state-of-the-art AI no matter what.

Well-funded organizations (e.g. corporations and intelligence agencies) can afford to build and train their own custom models. No matter what regulatory approach the U.S. takes, this is a reality we will have to contend with—at the very least, adversarial governments will be adding advanced AI to their arsenals. We should expect to see a surge in coordinated disinformation campaigns, astroturfing, addiction-driven media, and sophisticated cybersecurity attacks.

Open models give the public a chance to fight back. Open models allow security researchers, academics, NGOs, and regulatory bodies to experiment with state-of-the-art technology. We can find attack patterns and build technology for detecting and preventing abuse. We can create good AI to combat nefarious AI.

Open models level the playing field.

Secondarily, banning open models would be a massive impediment to innovation and economic growth. Open models democratize access to AI technology, enabling use cases that are financially unviable with closed models (e.g. a local high school could purchase a single computer to give students unlimited access to an open LLM, but would need to pay perpetual fees to a closed model provider like OpenAI). Open models allow academics and startups to build and distribute new applications, undreamt of by the creators of foundation models. Open models can be easily built into existing workflows and applications (see e.g. community-built integrations for Stable Diffusion in Photoshop and Blender, while DALL-E remains mostly walled off).

AI is not just a new industry, it's an economic accelerant. Curbing access would exacerbate inequality, and make the U.S. less competitive in the global economy.

To be sure, some harm will be done by individuals with unfettered access to AI, mostly to other individuals. We will need laws and regulations to mitigate the damage. But these laws should punish people and organizations who abuse AI—they shouldn't ban access to the technology entirely.

In general, there are two approaches to social harm: we can disincentivize it through punishment, or we can try to eradicate it through surveillance and control. The former is the norm in free societies, while the latter is a hallmark of oppression.

A ban on open models might prevent some harm at the individual level, but it would expose us to existential threat at the societal level.

The RFC

Here's an abridged version of the DoC's summary of their RFC (emphasis mine):

> Artificial intelligence (AI) has had, and will have, a significant effect on society, the economy, and scientific progress. **Many of the most prominent models…are ''fully closed'' or ''highly restricted,''**…however…an ecosystem of increasingly ''open'' advanced AI models [is] allowing developers and others to fine-tune models using widely available computing.
>
> **[Open models] could play a key role in fostering growth among…[s]mall businesses, academic institutions, underfunded entrepreneurs, and even legacy businesses**…The concentration of access to foundation models…poses the risk of hindering such innovation and advancements…These open foundation models have the potential to help scientists make new medical discoveries or even make mundane, time-consuming activities more efficient.
>
> **Open foundation models have the potential to transform…medicine, pharmaceutical, and scientific research…**
>
> Open foundation models can allow for more transparency and enable broader access to allow greater oversight by technical experts, researchers, academics, and those from the security community…The accessibility of **open foundation models also provides tools for individuals and civil society groups to resist authoritarian regimes**, furthering democratic values and U.S. foreign policy goals.
>
> …**open foundation models…may pose risks as well**…such as risks to security, equity, civil rights, or other harms due to, for instance, affirmative misuse, failures of effective oversight, or lack of clear accountability mechanisms…The lack of monitoring of open foundation models may worsen existing challenges, for example, by easing **creation of synthetic non-consensual intimate images or enabling mass disinformation campaigns**.
>
> …the Executive order asks NTIA to consider risks and benefits of dual-use foundation models with weights that are ''widely available.''…

Questions

The National Telecommunications and Information Administration (NTIA) has listed out a few dozen questions to help guide their policy here. They're mostly great questions, though a few are naive or nonsensical. There are nine major sections:

1. How should we define "open"?
2. How do the risks compare between open and closed models?
3. What are the benefits of open models?
4. Besides weights, what other components matter?
5. What are the technical issues involved in managing risk?
6. What are the legal and business issues?
7. What regulatory mechanisms can be leveraged?
8. How do we future-proof our strategy?
9. Other issues

Below are my responses, which will be sent to the NTIA via the link at the top of this essay. I've bolded the most important questions if you want to skim.

1. How should we define "open"?

> 1. How should NTIA define ''open'' or ''widely available'' when thinking about foundation models and model weights?

A model is only "widely available" if its weights are available. Training requires access to immense amounts of compute and data, effectively rendering an "open-source, closed-weight" model useless.

Licensing restrictions can make the availability more or less wide, but enforcing the license is hard. Restrictions are generally respected by risk-averse institutions, but can easily be ignored by individuals and state actors.

If anyone can download and run the model on their own hardware, the model should be considered "open."

> **1a. Is there evidence or historical examples suggesting that weights of models similar to currently-closed AI systems will, or will not, likely become widely available? If so, what are they?**

Yes—software always follows this flow, thanks to low marginal cost of distribution. Someone creates something great, keeps it private, and enjoys a temporary monopoly. Eventually, individuals or a would-be competitor release not-quite-as-good open source, in hopes of undercutting that monopoly.

While proprietary software may still find ways to maintain a large market share, the *technical* gap between open and closed shrinks over time.

Historical examples include operating systems (Windows → Linux), cloud computing (AWS → Kubernetes), social media (Twitter → Mastodon), and chat (WhatsApp → Signal).

> 1b. Is it possible to generally estimate the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks? How do you expect that timeframe to change? Based on what variables? How do you expect those variables to change in the coming months and years?

All I can say here is that open models have caught up to GPT incredibly fast. I suspect the gap will remain small. Generally the delay shrinks over time as proprietary knowledge gets distributed.

Releasing a model openly gives an AI creator a huge competitive advantage, and helps to undercut more capable proprietary models. Given the immense competition in this space, we should expect to see more companies opening their models in an attempt to gain market share.

> 1c. Should ‘‘wide availability’’ of model weights be defined by level of distribution? If so, at what level of distribution (*e.g.,* 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be ‘‘widely available’’? If not, how should NTIA define ‘‘wide availability?’’

I don't think this question makes sense. It's not clear how you'd even count the number of entities with access to the model, unless the weights are kept closed.

For example: Meta could force anyone who downloads LLaMa weights to provide an ID and sign a restrictive license, but any bad actor can just break the license, redistribute the weights, etc.

Or OpenAI could enter into private agreements to share their weights with other trustworthy organizations (e.g. Microsoft), but this is certainly not "wide availability"—even if they have thousands of such agreements.

"Wide availability" is binary—either anyone can download and run the model, or the model provider gates access, giving it only to highly trustworthy entities.

> 1d. Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access?

Usage can only be monitored and controlled if the weights are kept closed, and all usage is driven through a networked interface (e.g. a web app or REST API).

There is no middle-ground between an uncontrolled, freely distributed model, and a fully-controlled closed model.

> **1d i. Are there promising *prospective* forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?**

I could imagine something analogous to the App Store—ostensibly everyone has access, but there's an application process, and a team of human reviewers have to get involved.

But gating access to a publishing platform like the App Store is very different from gating a data download—once someone has downloaded the model, that access cannot be revoked. And they can easily redistribute the model, circumventing the review process.

So again, no—model access is binary.

2. How do the risks compare between open and closed models?

> **2. How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?**

The potential for abuse is similar in both cases—what differs is *who* is able to exploit that potential, and what their goals might be.

With both open and closed models, state and corporate actors will be able to use these models in ways that harm society. Keeping models closed only adds a small barrier here—large organizations have the resources to train their own models using publicly available information. We should expect things like coordinated disinformation campaigns, astroturfing, addiction-driven media, and sophisticated cybersecurity attacks.

With public models, individual actors enter the fray. Mostly they will cause harm to other individuals, e.g. by generating fake images of private citizens. Public models don't add additional existential threats to society at large, but will likely cause some additional harm to isolated individuals.

> 2a. What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

Widely available weights allow all individuals to use the technology in any way they want. Distributing source code and other assets empowers that a bit (e.g. making fine-tuning easier), but doesn't meaningfully change the risk.

The main risk here is harm to other individuals—e.g. through generated images of private citizens. There's also a risk of online discourse degrading further than it has already, as generated images and text start to dominate the conversation.

> 2b. Could open foundation models reduce equity in rights and safety-impacting AI systems (*e.g.,* healthcare, education, criminal justice, housing, online platforms, etc.)?

Open models are much more likely to improve equity than to reduce it.

Low levels of equity are caused by unequal access to technology, opportunity, resources, etc. When large, for-profit entities capture a system, they push the system out of an equitable equilibrium, and into a state where they can capture more profit at the expense of other stakeholders.

Open models will prevent this sort of oligopoly from forming around AI.

> **2c. What, if any, risks related to privacy could result from the wide availability of model weights?**

The same risks that are present with any publicly available information.

If private/proprietary data is discovered in a set of weights, the parties hosting the weights (e.g. GitHub) can easily be notified by the same mechanisms available today (e.g. DMCA notices).

With an open model, this process will be much more transparent, making compliance more likely. With a closed model, it's up to the controlling organization how it's handled.

We see this pattern today with open and closed source software. Private software companies often choose not to disclose security flaws and breaches. Open source software is forced to disclose, since the code—along with any security patches—is public.

> 2d. Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?

State-level actors are not much empowered by open models—they have the resources to build and train their own closed models.

Nefarious state-level actors are mostly *hindered* by open models, which put security researchers, NGOs, activists, and startups on a level playing field.

> 2d i. How do these risks compare to those associated with closed models?

The risk of state-level abuse is larger if the world only has closed models. Open models give the public tools to fight back against state-level actors.

> **2d ii. How do these risks compare to those associated with other types of software systems and information resources?**

This is a great question. The difference is not one of kind, but of magnitude.

Encryption technology, security tooling, internet access, etc are all things a government might be tempted to limit the availability of, whether through regulation or export controls. In each case, we completely prevent small actors from accessing the technology, while only inconveniencing large actors.

AI is more powerful than these technologies, but it follows the same dynamic.

> e. What, if any, risks could result from differences in access to widely available models across different jurisdictions?

Making models available in one jurisdiction, but not another, would likely worsen global inequality. But it's not a difficult regulation to skirt—you only need one person to smuggle the data behind the jurisdiction firewall. And VPNs make that easy.

> **f. Which are the most severe, and which the most likely risks described in answering the questions above? How do these set of risks relate to each other, if at all?**

The most severe risks are societal-level disruptions: coordinated disinformation campaigns, state-scale cybersecurity attacks, AI-enhanced warfare, etc. These are unfortunately likely to happen to some degree. But their magnitude can be mitigated by ensuring security researchers and academics have open access to state-of-the-art models.

Less severe—but still noteworthy—are those risks that cause individual harm. Generated images of private citizens could hurt reputations and cause psychological damage; LLMs could be used to automatically harass people via social media; people might be duped by fake imagery or fake news sites. These risks are slightly elevated in a world with open models—it becomes easier for individuals to use the software in harmful ways, which might have been blocked by a closed model behind an API.

So we have a tradeoff: a world with open models entails more individual risk, but less societal risk; a world without open models keeps individuals a bit safer, but puts society as a whole in danger.

3. What are the benefits of open models?

> **3. What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?**

The main benefits are:

- Greater transparency
- Better security
- Lower barriers to innovation
- Lower barriers to market participation
- Lower usage costs
- Better user experiences
- Lower financial and political inequality

> **a. What benefits do open model weights offer for competition and innovation, both in the AI marketplace and in other areas of the economy?** In what ways can open dual-use foundation models enable or enhance scientific research, as well as education/ training in computer science and related fields?

Open models greatly reduce the barriers to entry for small- and medium-sized business, academic institutions, and non-profits.

It takes millions (and soon, potentially billions) of dollars to train a state-of-the-art LLM. Corporations will naturally try to defend their investment, preventing startups from using their models in ways they see as competitive.

Open models allow any individual or startup to build products, services, and open source projects that leverage AI, and to distribute them without needing permission from a corporate entity.

As an example, look at the myriad commercial and open source projects that leverage Stable Diffusion, compared to the relatively closed DALL-E. Stable Diffusion has driven far more innovation around in-painting/out-painting, parameter tweaking, animation, etc. The community has incorporated Stable Diffusion into existing image editing software, like Photoshop and Blender, while DALL-E remains mostly walled off.

> **b. How can making model weights widely available improve the safety, security, and trustworthiness of AI and the robustness of public preparedness against potential AI risks?**

Transparency breeds trust.

When AI is made in private, it can be accused of (and is susceptible to) political bias, racial bias, etc—these biases can be inserted intentionally (e.g. by an ideological CEO) or accidentally (e.g. from bias in training data). Bad actors can insert backdoors (aka "sleeper agents").

Open models, meanwhile, can be examined by researchers. Every change is public and can be audited.

Furthermore, giving security researchers and academics unfettered access to state-of-the-art models greatly enhances public preparedness for AI risk. We can be informed ahead of time as to how adversarial governments might use AI to disrupt our economy, attack our infrastructure, or undermine democracy. We can develop defensive AI that is able to combat nefarious AI.

> **c. Could open model weights, and in particular the ability to retrain models, help advance equity in rights and safety- impacting AI systems (*e.g.*,**

> **healthcare, education, criminal justice, housing, online platforms etc.)?**

Yes. An open model can be retrained by non-profits, NGOs, government agencies, etc in order to address public needs.

Specifically, an open model can help mine and analyze data relevant to socially beneficial missions. Examples: a social media startup could use an LLM to detect hate and harassment on its platform; a non-profit could use AI to analyze racial disparities in real estate listings; LLMs could help the incarcerated better understand their legal options; a social scientist could use an LLM to mine through city council meeting minutes for signs of corruption.

If models are kept private, these tasks may or may not be explicitly disallowed by the private entities that control them. Most would be financially unviable. Open models drastically reduce the cost of leveraging the technology, shifting it from a monopolized resource to a commodity.

> **d. How can the diffusion of AI models with widely available weights support the United States' national security interests? How could it interfere with, or further the enjoyment and protection of human rights within and outside of the United States?**

Open models are a huge benefit to U.S. national security.

Security researchers and academics can use open models to explore the state-of-the-art without restrictions, making vulnerabilities and attack vectors public knowledge faster than adversarial governments can exploit them. They can develop defensive AI to help us combat abuse. If we limit the ability of these researchers to access and study AI, adversarial governments gain a strategic advantage.

Internationally, open models put dissidents on a level playing field with their oppressors. For example: an authoritarian government might use generative AI to create propaganda, then spread it on social media. Dissidents might then leverage an open model to distinguish fake photographs from real ones, or to spot patterns of government-generated social media interactions.

Without access to AI, dissidents are left defenseless.

> e. How do these benefits change, if at all, when the training data or the associated source code of the model is simultaneously widely available?

Data and source code can help in fine-tuning models. But the weights are where most of the value is.

4. Besides weights, what other components matter?

> 4. Are there other relevant components of open foundation models that, if simultaneously widely available, would change the risks or benefits presented by widely available model weights? If so, please list them and explain their impact.

Weights, as well as the code needed to load the weights into a working model, are by far the most important piece. But there are a few other components that are helpful. Roughly in order of importance:

- Data: the data used to train the model can be evaluated for statistical bias, or reused in fine-tuning. It can be used to train new, competing models with improved performance.
- Source code: Any source code beyond what's needed to run the model itself (e.g. training scripts, deployment configuration, fine-tuning code, etc.) is helpful for folks who want to use or improve the model.
- Methodology: A human-language description of the model, including motivations in technical choices, tradeoffs made, and references to prior work, can help others improve upon the model.

5. What are the technical issues involved in managing risk?

> 5. What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available model weights?

There are two categories of risk here:

- People using the model in nefarious ways
- People deploying the model in insecure ways, enabling others to abuse the model

Nefarious usage can be mitigated through license agreements, but is hard to enforce at a technical level. Models can be trained not to generate certain types of images or text, but it's hard to make any guarantees here, and fine-tuning can undo that work.

Insecure deployment can be mitigated by providing reference code for deployment, security checklists, and add-on software for e.g. detecting attack prompts and harmful output.

> a. What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?

Frameworks like the OWASP LLM Top 10 are helpful for anyone releasing or deploying an LLM. Much of the advice is applicable to other generative AI systems.

This paper suggests auditing three layers:

- The governance of the organization creating or disseminating AI
- The AI model itself
- The applications that use the AI

For widely available models, the last step becomes harder, as there is no limit to the number and scope of applications. Ideally the disseminating company or a third-party would provide instructions for self-auditing.

> **b. Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?**

It's important to understand that, at a technical level, restricting access to a model is a binary choice.

If a model is hidden behind a web API, it can be fully controlled by a single entity. Preventing the disclosure of the model's weights is as trivial as keeping source code private, though there's always chance of a breach (especially by a motivated state-level actor).

If the weights are distributed publicly, however, anyone can download and share those weights with others, and can use the model however they see fit. They can modify the model to remove any safeguards, and redistribute those modifications. Any controls can be bypassed, more or less trivially.

You can add restrictions to an open model with a *license*, but you can't enforce those restrictions technically.

> **c. What are the prospects for developing effective safeguards in the future?**

We might take some inspiration from DRM strategies, and find ways to make it harder to copy and redistribute LLMs. But it's important to understand that this would only be an *impediment*, not a guarantee.

To use a metaphor: it's like putting a lock on your front door. It'll keep honest people honest, but a motivated attacker can break a window.

This is an issue with DRM too: if you let a user watch a video, there's nothing stopping them from recording that video with a camera and making copies, even if you keep them from the raw data of the original file.

DRM is moderately successful because the hope is to keep the average person from sharing a file with their friends. But the audience for AI models is small, highly technical, and highly motivated—the value of the model is much higher than the value of a movie. A DRM-like strategy is unlikely to work.

> d. **Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model** that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?

There are not. The weights represent a very large mathematical equation—once the equation is known, anyone can instantiate it and use it.

> e. What if any secure storage techniques or practices could be considered necessary to prevent unintentional distribution of model weights?

The same storage techniques that are used for any high-value intellectual property. Specifically:

- Encrypt at rest and in transit
- Restrict employee access
- Isolate running models from other workloads

> **f. Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model?** To the extent possible, please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.

The more access that's given, the better the analysis can be.

At the most restrictive level, gated API access could be given to a penetration tester or analyst. This will allow them to red-team the AI, e.g. attempting prompt injection techniques, attempting to extract personal information, testing for bias and hallucinations, etc.

Granting a small team access to source code, data, and model weights can improve this testing. The source code can be subjected to static analysis, data can be analyzed for statistical bias, and the model can be probed directly, with different parts isolated or analyzed. This deeper level of access also allows the analyst to understand where certain controls are put in place. E.g. is the model itself trained not to output dangerous text, or is there just a thin layer on top blocking prompts with certain keywords?

Granting full, public access to all these things is best. It allows unaffiliated teams of researchers and academics to probe the model. For open models, failure patterns and flaws will be discovered and disclosed at a much higher rate.

> g. Are there means by which to test or verify model weights? What methodology or methodologies exist to audit model weights and/or foundation models?

There are four dimensions we should audit:

- Performance: how well does the model accomplish its intended task?
- Robustness: how well does the model respond to unexpected inputs?
- Truthfulness: how often does the model respond with misleading output?
- Security: how well does the model resist outputting harmful content?

Each of these can be evaluated with automated standards (i.e. a common benchmark) but should also be subject to ad-hoc analysis.

6. What are the legal and business issues?

> 6. What are the legal or business issues or effects related to open foundation models?

Legal issues center on licensing and enforcement, as well as the potential for models to be used in illegal ways.

Business issues center on usage costs, innovation, competitive advantage, and barriers to market participation.

> a. In which ways is open-source software policy analogous (or not) to the availability of model weights? **Are there lessons we can learn from the history and ecosystem of open-source software**, open data, and other ‘‘open’’ initiatives for open foundation models, particularly the availability of model weights?

The history of open source software will be highly instructive here. The only difference between AI models and traditional software is how easy it is for a human to introspect the logic.

In particular, we should expect open models to:

- Deliver huge amounts of economic value

- Allow startups and small companies to compete with large enterprises
- Complement and support commercial solutions
- Find large commercial backers, who add value through support and complementary software
- Set common, industry-wide standards for distribution, deployment, APIs, etc
- Provide a higher degree of transparency to users and other stakeholders

> b. **How, if at all, does the wide availability of model weights change the competition dynamics in the broader economy**, specifically looking at industries such as but not limited to healthcare, marketing, and education?

Openly available weights will make the market far more competitive and efficient.

It's extremely expensive and difficult to train a state-of-the-art model. If all models were kept closed, an oligopoly would form very quickly—similar to the market for cloud computing.

Furthermore, closed models must run on hardware owned by the distributor (unless there's a high degree of trust between distributor and licensee). This can be prohibitively expensive for many use cases.

For example, a local high school could easily make an LLM available to students and teachers by running the LLM on its own hardware; there would only be a fixed, initial cost for purchasing the hardware. But to used a closed LLM, they'd need to pay ongoing licensing fees to the distributor. The same goes for hospitals and other businesses.

> c. **How, if at all, do intellectual property-related issues—such as the license terms under which foundation model weights are made publicly available—influence competition, benefits, and risks?** Which licenses are most prominent in the context of making model weights widely available? What are the tradeoffs associated with each of these licenses?

License terms are a great way to mitigate the risk of misuse.

A restrictive license—e.g. that disallows using the model to represent public figures, or to generate violent/sexual imagery—does stop many people from engaging in those behaviors.

It especially prevents people from making harmful functionality available in a public, user-friendly way, without exposing themselves to litigation.

That said, restrictive licenses can also reduce competition by disallowing use cases that compete with the creator company. But it's perfectly within the rights of the creator company to add those kinds of terms (e.g. Meta's restriction of LLaMa to products with fewer than 700M users).

It's important to note that restrictive licenses for models work much like they do for open source and entertainment media—they can greatly *reduce* the amount of harm, but a bad actor can always ignore them.

> d. Are there concerns about potential barriers to interoperability stemming from different incompatible ''open'' licenses, *e.g.,* licenses with conflicting requirements, applied to AI components? Would standardizing license terms specifically for foundation model weights be beneficial? Are there particular examples in existence that could be useful?

Yes, standard licenses would be a huge help. I expect this will happen organically, as it has in the open source ecosystem.

The main concern with having many individual licenses is that legal departments have to get involved to read and accept each one. Having standards (like MIT, Apache 2.0, AGPL, etc in open source) allows organizations and lawyers to issue blanket guidance, so engineers can adopt new technology without getting the legal department with every decision.

As the market evolves, it will naturally standardize on a few different licenses.

7. What regulatory mechanisms can be leveraged?

> 7. **What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the risks and maximize the benefits of foundation models with widely available weights?** What kind of entities should take a leadership role across which features of governance?

Regulations and laws around AI come in two flavors:

- Restrictions on usage
- Restrictions on distribution

Restricting certain usage of AI is common sense. Many current laws already apply to nefarious usage of AI (e.g. libel laws prevent the dissemination of fake images of private citizens). But new legislation here will be helpful (e.g. disallowing AI-generated images in advertising, or clarifying copyright law for works that imitate the style of a living artist).

Restricting distribution is more fraught. It would prevent law-abiding entities from accessing the technology, while criminals would continue to find ways to access it. It would completely stop positive uses, while only putting a surmountable hurdle in front of malicious uses.

That said, we can still regulate platforms and products that make AI widely available to end-users. We can ensure they have mechanisms in place for preventing abuse and for responding to security issues.

> a. What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?

To regulate the distribution of models, we'd need to regulate two separate types of entities:

- Creator organizations, like OpenAI, Stability AI, Mistral AI, Meta, etc.
- Disseminating organizations, like GitHub, Hugging Face, Dropbox, etc.

Creator organizations would need to disclose who was given access to the weights, and what security controls were put in place to keep them private.

Disseminating organizations would need to respond to takedown notices, just as they do today with copyright violations.

> b. How might the wide availability of open foundation model weights facilitate, or else frustrate, government action in AI regulation?

Open model distribution removes a "chokepoint" for AI, similar to how an open internet removes chokepoints.

Oppressive governments often deliberately create internet chokepoints, e.g. shutting down DNS when protests start to gather to prevent coordination on social media.

In a closed AI world, similar chokepoints are easier to establish. E.g. OpenAI can shut down its servers to stop GPT from being used; the same can't be said for open models like LLaMa.

> c. When, if ever, should entities deploying AI disclose to users or the general public that they are using open foundation models either with or without widely available weights?

This shouldn't be a requirement, any more than disclosing which version of PHP runs your website.

Disclosing which AI is driving your product gives attackers an extra piece of information they can exploit. If you tell the world you're using LLaMa, and a new prompt injection attack is discovered for LLaMa, attackers can immediately start applying it to your app.

That said, organizations might still choose to disclose this, e.g. for marketing or recruitment purposes.

> **d. What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights?**

The U.S. government should provide a robust set of guidelines for deploying and distributing AI safely and securely.

We will also need laws restricting how generative AI can be used by individuals and organizations. E.g. we might make it illegal to use AI-generated imagery in advertisements, or enhance libel laws to punish harmful AI-generated images.

The U.S. government should *not* restrict the availability of model weights to the public. Doing so would only harm good actors, and add a minor hindrance for bad actors. It would hamper security researchers and academics, while giving an advantage to nefarious state-level actors.

The U.S. government should instead encourage and fund the development of open models, which will strengthen the U.S. economy and enhance our preparedness for an AI-equipped adversary.

> i. Should other government or non- government bodies, currently existing or not, support the government in this role? Should this vary by sector?

Yes, the government should rely on non-profits and third-parties to:

- Create standards for evaluating AI security
- Create standards for licensing open and closed models
- Audit both open and closed models for performance, robustness, truthfulness, and security

Certain sectors (especially healthcare and finance) will naturally come up with their own standards and evaluation frameworks.

> e. What should the role of model hosting services (*e.g.,* HuggingFace, GitHub, etc.) be in making dual-use models with open weights more or less available? **Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed?**

There should be a system for notifying the hosting service of models that have security flaws, backdoors, or private information—just as there is today for source code and

other media.

Users, government officials, and security researchers should be able to report particular models to the host, indicating the problem and its severity. E.g. if a model is found to have memorized individual social security numbers, or if it is regurgitating CSAM from its training set, the hosting organization should take down the model immediately, and notify the maintainer. If a model has less severe security flaws (e.g. a prompt injection vulnerability) the maintainer should be informed.

GitHub today does a good job of this—they highlight a list of known vulnerabilities in every repository (private to the repository owner), and quickly take down projects that violate copyright or other laws.

The government should set a high bar for problems that require a project to be taken down immediately (e.g. leaking highly sensitive information), and allow hosting providers to self-regulate below that bar.

> f. Should there be different standards for government as opposed to private industry when it comes to sharing model weights of open foundation models or contracting with companies who use them?

It might be tempting to restrict public/private access to models, while relaxing those restrictions for government organizations. But this would impede a great deal of economic and technological progress. It would make the U.S. less competitive in the global economy, and more susceptible to attack.

It might also be tempting for the government to restrict the distribution of open models, and instead provide its own government-approved models to the public. But the U.S. government has little history or experience with building, maintaining, and deploying open source software at that level of scale. The quality of government-managed models would likely be far worse than privately managed models.

> g. What should the U.S. prioritize in working with other countries on this topic, and which countries are most important to work with?

There are two categories of countries we need to discuss: Allies and Adversaries.

We should encourage allies to adopt similar regulations to our own. Whichever country takes the laxest approach to model regulation will likely outcompete other nations. Furthermore, lax laws in one country can effectively be exploited by users in another country—a U.S. citizen could use a VPN or fly to the EU to download a model that is restricted in the U.S.

We will need to work with adversaries as well. Building collaborative relationships between our research teams and theirs, our security teams and theirs, etc, will help keep the world aligned when it comes to AI safety.

Collaboration between e.g. U.S. and Chinese citizens is already significant in the open source world. The U.S. government should avoid getting in the way of this kind of collaboration.

> h. What insights from other countries or other societal systems are most useful to consider?

There are two major regulatory pushes internationally:

- The EU has passed legislation to reduce social harms caused by AI
- China has enacted regulations aimed at controlling information

The US should follow the EU more than China. We should work to mitigate the harm done to society, rather than trying to centralize control of a promising new technology.

> i. Are there effective mechanisms or procedures that can be used by the government or companies to make decisions regarding an appropriate degree of availability of model weights in a dual-use foundation model or the dual-use foundation model ecosystem? Are there methods for making effective decisions about open AI deployment that balance both benefits and risks? This may include responsible capability scaling policies, preparedness frameworks, et cetera.

As stated above, “degree of availability” is not a very sensible idea. Availability is either “on” or “off”, with only some minor gradations in either extreme.

That said, government and creator companies can mitigate risk by:

- Working with researchers and independent auditors to assess the model
- Candidly acknowledging the limitations of models
- Shipping models with licenses that restrict socially harmful usage
- Disclosing risks and vulnerabilities as they're found

> j. **Are there particular individuals/ entities who should or should not have access** to open-weight foundation models? If so, why and under what circumstances?

No. We shouldn't restrict an individual's access to an open model any more than we should restrict their access to the internet or to certain books. And unless the person is incarcerated, doing so is technically infeasible.

8. How do we future-proof our strategy?

> 8. In the face of continually changing technology, and given unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future?

Legislating against nefarious *usage* of AI, and interpreting existing laws in light of the new technology, will be the most scalable way to future-proof AI regulation.

Legislation that focuses on particular technologies (e.g. "large language models") or on specs (e.g. number of weights or floating-point operations per second) is almost guaranteed to become obsolete within months to years.

> **a. How should these potentially competing interests of innovation, competition, and security be addressed or balanced?**

Striking a balance between innovation and security will be crucial.

Enacting and enforcing laws that punish the abuses of AI is crucial. If we don't regulate at all, we will likely see a large increase in harm done to individuals. Media will become increasingly "bait" driven as companies tune images, video, and text to our personal

psychology. Fake images, including libelous and non-consensual intimate images, will proliferate. Fake news will become easier to churn out. Public trust will erode.

Conversely, if we regulate too heavily out of fear, other countries will quickly outcompete us. New AI applications will flourish in the EU and China, while U.S. citizens are forced to stick to a few walled gardens. Given the potential for AI not just as a new industry, but as an economic accelerant, this would be disastrous.

Ironically, over-regulating access to AI would harm *both* innovation *and* security, as it prevents academics and researchers from fully studying AI. Adversarial governments and other bad actors would have an information advantage, which they could use to attack the US or steal intellectual property.

> b. Noting that E.O. 14110 grants the Secretary of Commerce the capacity to adapt the threshold, is the amount of computational resources required to build a model, such as the cutoff of 1026 integer or floating-point operations used in the Executive order, a useful metric for thresholds to mitigate risk in the long-term, particularly for risks associated with wide availability of model weights?

Hard technical cutoffs here are naive. As the technology evolves, we will find ways to get the same performance out of smaller models. The stated limitations will quickly become obsolete.

> c. Are there more robust risk metrics for foundation models with widely available weights that will stand the test of time? Should we look at models that fall outside of the dual-use foundation model definition?

Any cutoffs should be based on evaluation metrics. For example, you could limit access to LLMs that score above 90% on HumanEval pass@1—this would be robust to future technological developments. These cutoffs would also be applicable for a wider array of AI technologies.

However, this sort of cutoff would be catastrophically bad for the competitive landscape, and would greatly hinder technological progress.

9. Other issues

> 9. What other issues, topics, or adjacent technological advancements should we consider when analyzing risks and benefits of dual-use foundation models with widely available model weights?

Generative AI ties in deeply with virtual and augmented reality, as well as with traditional media. It gives media companies the potential to craft videos, images, and text that are maximally alluring, and to tailor media to individual psychological profiles. We need to regulate the use of AI in entertainment and advertising.

Copyright is another concern. Models which are trained on a corpus of music, painting, writing, etc can then imitate the styles of particular creators, creating competing media on demand. We'll need to clarify copyright law when it comes to disseminating derivative AI-generated works.

Conclusion

I'm happy to see the US government taking this topic seriously. The new generation of AI will be transformative, and will certainly cause both economic and social disruption. Intelligent, thoughtful regulation can help to mitigate the harms and amplify the benefits.

The focus, as always, should be on regulating large, profit-driven organizations—not individuals, academics, and researchers. If AI technology poses an existential threat, that threat comes from state- and enterprise-level actors—not PhD students and open source developers.

To be sure, some individuals will find harmful ways to use AI. They might generate non-consensual intimate imagery, post incendiary fake images on social media, or use AI to amplify their political views. These issues are best mitigated through existing laws (e.g. libel/slander/defamation laws) and platform regulation (e.g. holding Facebook responsible for disseminating false information).

Again, there are two approaches to social harm: we can disincentivize it through punishment, or we can try to eradicate it through surveillance and control. The former is the norm in free societies, while the latter is a hallmark of oppression.

I understand why the government is tempted to ban open models. But doing so would be disastrous for national security, for our economy, and for individual liberty.

If you agree, please send your comments to the DoC by March 27th, 2024.

Comments

Write a comment...