My name is Zvi Mowshowitz. I write the substack Don’t Worry About the Vase, where every week I collect and discuss recent developments in AI, covering everything from mundane utility to future existential risks.

Opening up the weights of an AI model is an irreversible act of proliferation. Once this occurs, anyone in the world has the power to, within a day and at trivial economic cost, disable any safety protocols or behavioral controls upon that model, and to hook that AI up to additional amplifying tools, unleashing its full underlying capabilities for any and all purposes. Many eagerly do exactly this. We in many ways do not know, even years later, what the full capabilities of our existing AI systems will be if given access to future tools.

Any such model will then be, in its fully unlocked form, in the hands of every corporation and individual, every state and non-state actor. Many will seek to further enhance its capabilities, and attempt to misuse it to do harm, or to make it as autonomous as possible.

If we share the weights for models at the frontier:

* We forfeit our lead in AI. China catches up.
* We render it impossible for labs to take safety seriously while remaining competitive.
* We do not know the full capabilities of such models, when enhanced with future tools.
* If we realize we made a mistake in releasing a model, we cannot undo that mistake.
* If harm does occur, there is no way to hold anyone properly legally responsible for it.

Thus my weekly column has a recurring section: Open model weights are unsafe and nothing can fix this.

That does not mean open model weights have no place. When models that lack dangerous capabilities are shared freely, this has a multitude of benefits, including to AI research and safety research. We should welcome open model weights for such systems such as Llama-2. The worry is that this may, if unconstrained, carry over to future models as strong and then stronger than GPT-4.

This would then pose increasingly severe catastrophic or existential risks from misuse, competitive dynamics or loss of control. It would irreversibly render all government policies and interventions increasingly difficult, expensive and intrusive to implement, whatever the policy goal might be.

Even if one’s goal was as limited as protecting copyright or preventing the creation of deepfakes or child pornography, preventing cyberattacks or having it not provide instructions to create a bioweapon, there would be no practical way to do this short of monitoring all individual computers. Open source AI image models are a clear illustration of this.

Open model weights also do not offer many of the benefits of other open source software. Open source software is typically more secure because anyone can examine the source code to find and fix flaws. Alas, open model weights does not mean open source code.

* The weights are an inscrutable black box.
* The method of generating those weights is rarely shared alongside the box.
* This does not allow the same collaborations, transparency or robustness benefits. The world cannot collaborate on methods or code in the typical open source way.

There is also a problem of incentives. Almost everyone working on something like Linux or Python wants the system or language to be secure. An unsafe operating system like Linux would be dangerous to its user. With AI models, many users want a fully unrestricted model, and an ‘unsafe’ or unrestricted system becomes dangerous to others more than to its user.

The executive order correctly identifies training compute thresholds as our best available tool to distinguish potentially dangerous models we must carefully monitor from smaller models where a free approach is safe and offers benefits.

Our policy on open model weights should follow the same principle.

Above the reporting threshold of 10^26 training flops, we should require strict cybersecurity measures to secure model weights, so they are not stolen by China or other adversarial state or non-state actors. Securing the weights necessarily entails not allowing them to be openly released. At a minimum, even if we for now leave such security to corporate discretion, it is highly unsafe and against the interests of national security to allow weights of such a model to be released.

Well below that threshold, we should allow or perhaps encourage the opening up of model weights. But we should also clarify and ensure that opening up the model weights in a negligent way makes you legally liable for any resulting harm, and strongly consider requiring the purchase of robust insurance against this above some intermediate threshold, the same way we require insurance to drive a car. We should then continuously revisit these rules and thresholds, and adjust them as we learn more.