



~ Contra Zuckerberg on ‘Open Source AI’

Friday, July 26th, 2024

In [internet tradition](#), I am putting aside my other duties and taking the day to write.¹

Clearly Mark Zuckerberg’s letter, [“Open Source AI Is the Path Forward”](#), accompanying the exciting release of Llama 3.1, was not intended to withstand intellectual critique. It was meant to rally goodwill and support for Meta AI from AI developers, policy-makers, and society by appealing to principles of openness and software freedom.

I strongly believe in software freedom, and I’m no fan of OpenAI’s ‘Closed AI’ approach. However, there are glaring flaws in Zuckerberg’s framing of his vision as ‘Open Source AI’ and his analogy with the history of Linux, and there are important gaps in his argument for the safety of pursuing this path towards a future with advanced AI systems.

If you haven’t read Zuckerberg’s letter, I honestly can’t say I recommend you read it (or, by extension, this critique). Instead, please consider spending some time reading something coherent such as [Vitalik Buterin’s vision for the future of AI](#). But if you did read Zuckerberg’s letter, I feel compelled to offer you a spoonful of medicine to stop it going down.

§ A summary of Zuckerberg’s letter

(You can skip this if you remember the letter.)

Zuckerberg’s letter begins by telling the story of the ascent of the Linux family of operating systems from underdog in a time of closed-source operating systems to its modern supremacy in server-side and mobile computing. He draws an analogy to the modern battle between closed AI models like those from OpenAI and comparatively open models like Meta AI’s Llama herd. Zuckerberg lays out a vision for the future of AI in which open models like Llama are an industry-standard foundational platform for an ecosystem of tools and data sets for tuning, distillation, and prompting models, not reliant on any one closed vendor’s AI API.

The bulk of the letter is a list of arguments for why this approach, dubbed ‘Open Source AI’, is the best path forward for AI application developers, for Meta, and for the world.

- For developers, Zuckerberg argues that developers gain by avoiding dependence on a closed AI vendor, namely in terms of freedom to tune and distil custom models, cost saving, protection against changing models or terms of service, and keeping data in-house. Moreover, he bets that open models will become fundamentally better than closed models in the future and so it’s worth investing in the open platform long term.

Later, Zuckerberg recounts his struggles with the restrictions placed on Facebook apps by Apple’s iOS platform, reinforcing the importance of freedom from platform constraints for businesses.

- Zuckerberg then outlines how Meta AI itself benefits from other people's contributions to the AI ecosystem built atop its models, how he doesn't see it possible to retain a competitive advantage in AI models in the long term anyway, and how AI access is not core to Meta's business model.² He also points out that Meta has a long history of releasing open source developer tools like PyTorch and React.
- For the world, Zuckerberg outlines the potential societal benefits of AI. Zuckerberg recognises that AI also brings risks of harm, and he separates these into (1) unintentional harms, from subtle harms through manifold daily user interactions to catastrophic harms from loss of control of powerful systems, and (2) intentional harms by bad actors, with either small resources or state-level resources. In each case, he attempts to argue that the world is safer under his 'Open Source AI' path than a closed alternative.

§ Zuckerberg's vision is not 'Open Source AI'

I am struck by several disanalogies between the ascension of Linux and Zuckerberg's vision for 'Open Source AI', taking the Llama 3.1 release as a prototype. Ultimately, I don't think Zuckerberg's vision is open enough to achieve the benefits he promises.

Where is the source? Most prominently, Llama 3.1 is what I would call an 'open weights' release rather than an 'open source' release, and this difference is crucial to some of the claimed benefits of 'Open Source AI'.

As far as I can tell, what has been released is the following:

- The weights of the trained and tuned models, available for download from Meta (or other hosts) after signing a licence agreement.
- A [GitHub repository](#) containing documentation and utilities for running the models.
- A [self-published Meta AI report](#) containing moderately detailed information on the model's architecture, training, tuning, and evaluation.

The details provided in the report are apparently more generous than previous similar details released by closed AI vendors. However, they are not enough for users to easily modify the training process and build their own version of the models (this being one of the capabilities Zuckerberg recognised as instrumental in the early stages of Linux's history).

I would argue that the correct analogy for source code for an LLM is not the weights and architecture, but (1) the software used for training and tuning the model (and that software's source code in turn), and (2) the data sets used during training and tuning (and the software used to collect it, and that software's source code in turn).³

So far, these tools and data sets have not been forthcoming.

Where is the freedom? Even if future releases are accompanied by source code and data, this may not be enough to realise the benefits Zuckerberg is appealing to. It also depends on what people are allowed to do with the model.

If the goal is freedom (as in Zuckerberg's appeal to the tyranny of Apple over its iOS platform), then there is more that one can insist on than mere access to the source code—you also want a permissive licence that gives you the freedom to do what you want with the software.

The Llama 3.1 herd is released under a custom license called the [Llama 3.1 community license agreement](#). I'm not a lawyer but a couple of the clauses stood out to me while reading the license that seem not exactly consistent with software freedom.

- Your use must comply with applicable laws and regulations (fine, I respect most laws and regulations...) and adhere to the [Llama 3.1 Acceptable Use Policy](#) (ah—there it is).

Skimming this policy, many of the prohibitions seem defensible—no human trafficking, no sexual violence, no weapons development, etc. Other things seem understandable, but overly broad and restrictive—no operation of transportation technologies? Moreover, I’m instantly worried that (1) in any subjective cases, as the publisher of this policy, Meta is the one who decides how the subjectivity is resolved, and (2) can this policy, referenced by URL from within the licence agreement, be updated by Meta at any future time?

- If you ‘create, train, fine tune, or otherwise improve an AI model’ and then release it, you must name the model something starting with ‘Llama’. For other products, you have to acknowledge prominently that the product was built with Llama.

On the one hand, so what, but on the other hand, how draconian to retain *naming rights* over derivative models?

- If at the time of the Llama 3.1 release your organisation caters to over 700 million monthly active users, this license is not available to you and you need to seek a custom license from Meta.

This seems targeted at existing serious competitors, whom I’m not going to stand up and defend. It’s nice that they can’t use it to cut down new competitors in the future since it’s pinned to the release date, except that a similar clause in the next generation of base models will just reset the cut-off date.

- You lose this license if you sue Meta because some model’s “outputs or results constitutes infringement of intellectual property or other rights owned or licensable by you”.

This one is a little on the nose, given that one can speculate that the main motivation behind the decision *not* to release training data is that Meta’s right to train AI models on that data is not clearly established.

Particularly that first restriction seems to cut pretty directly against, say, Richard Stallman’s [freedom 0](#). By contrast, Linux is licensed under the GPL. I think there is legitimate debate in the community about what one should be allowed to do with an AI model (or any piece of software), but it’s notable that with this license Meta is retaining the right to have the final say in that debate.

Is the freedom exercisable? Even if a future release came with a much more permissive license—what good is the freedom to modify a piece of software if only Big Tech companies have the resources to do so?

I’m not familiar with the computing resources required to compile one’s own Linux kernel back in the day, but training the Llama herd reportedly took “39.3M GPU hours of computation on H100-80GB (TDP of 700W) type hardware”. That is a lot of compute—totally out of reach for individuals and all but the largest organisations.

At the moment the amount of computing resources required to train frontier AI models is a fundamental barrier to *realising* an important software freedom, meaning even if source code were released, perhaps these models would be free-in-name-only.

It’s actually somewhat hard to imagine surpassing this problem even in principle. Maybe as the field of AI advances, the cost of training models at a given scale will continue to drop, and resources available to individuals will continue to grow exponentially.

However, training models *at the frontier of scale and performance* might always cost resources that are, by definition, exclusive to those that can pay the most. So, might truly

free frontier models be fundamentally impossible? I think this is an open question in AI software freedom.

What would we do without Meta? Zuckerberg doesn't propose that individual hackers or businesses want to or need to be free to train their own Llama models from scratch—I'm the one insisting on that absolute standard. But relaxing this standard and settling for a 'free from the weights up' model is problematic, in that it doesn't remove a crucial dependence on Meta itself.

I admit that an open weights release is partially open and does confer some freedom to modify the software. Namely, they can (feasibly!) modify the software to the extent achievable by tuning, distillation, and prompting.

The question is, is this enough freedom? I don't think so. There seem to be some modifications that we might want to make that these methods can't achieve—only access to the 'true source' (tools, data sets, and resources used for pre-training) can achieve them.

1. One important example is that for certain dangerous capabilities in the base model, 'safety tuning' methods (so far) have limited effectiveness in preventing jailbreaks from exposing these dangerous capabilities in adversarial contexts.
2. A second example is that it appears reaching progressive generations of base models themselves requires interventions at the pre-training stage. You can't get from Llama 2 to Llama 3.1 by fine-tuning.

In the theme of disanalogies to [the history of Linux](#), Zuckerberg is proposing an ecosystem where anyone in the community can participate in a surface-level "Bazaar" and contribute tuning data sets and methods for patching shallow issues, but fundamental improvements can only come down from the "Cathedral" on high.

This leaves the community with a crucial dependency on Meta itself. If, as Zuckerberg points out, "organisations don't want closed model providers to be able to change their model, alter their terms of use, or even stop serving them entirely", why would they be any safer placing their trust in Meta? It seems like Meta can still do all of these things to developers.

I'd put more faith in the long-term competitiveness of the Llama platform if the training tools and data sets were released under an open license so that at least the community *only* has to find a ridiculous amount of computing resources to continue the Llama line in the event that Zuckerberg's mood changes.⁴

Llama 3.1 is undoubtedly an impressive technical accomplishment. But it's important to realise that it's an accomplishment of the engineers and researchers at Meta AI, not of the open source community.

§ **Zuckerberg's argument for safety is insufficient**

Switching to another part of Zuckerberg's message—Zuckerberg attempts to argue that 'Open Source AI Is Good for the World'. We can certainly agree on the potential benefits of AI. I also want to give Zuckerberg credit for taking potential harms seriously enough to address them in the letter.

Unfortunately, we appear to disagree on the extent to which the potential harms are successfully addressed by the 'Open Source AI' vision. I found Zuckerberg's argument lacking. I'll follow Zuckerberg in dividing the potential harms into unintentional harms and intentional harms, and then comment on the gap between relative safety against these classes of harms and AI that is beneficial to the world.

Unintentional harms. I think Zuckerberg’s argument here can be distilled down to the following sentence:

“ Open source should be significantly safer [than closed AI systems] since the systems are more transparent and can be widely scrutinized.

This is essentially an appeal to [Linus’s law](#) of “Bazaar”-style open source software development: the observation that “given enough eyeballs, all bugs are shallow.”

Zuckerberg’s plan does appear to benefit from this effect to a limited extent (and perhaps more-so than closed AI systems). Since the weights are open anyone in the ecosystem can perform their own additional safety testing, and if a model anywhere in the ecosystem shows unintended behavioural issues this can be recognised and ‘patched’ by fine-tuning or prompting and this patch can be propagated through the ecosystem to other developers using the same base model.

I don’t agree that this leads to a significant improvement in safety for the harms Zuckerberg cites as examples.

1. For harms arising from “what influence AI systems will have on the billions of people who will use them”, it’s important to consider even [diffuse, communal harms accrued over a long time](#), which when aggregated over billions of people still amount to a major problem. LLM behaviour is stochastic and context-dependent, and so these subtle harms might take a long time to detect, despite the large number of eyeballs at play.

Even if these harms are detected, they will be difficult to attribute to a particular cause and they will be difficult to fix. There are many disparate influences on the AI system’s behaviour, from Meta AI’s pre-training and tuning data to a patchwork of further tuning data assembled from the ‘Open Source AI’ ecosystem to the system prompts provided in the final application, for example. Meta’s data sets are not open to public scrutiny and I don’t suppose application developers will necessarily open their own tuning data or system-prompting practices to scrutiny either. Moreover if the issue is fundamental enough to have come from the “Cathedral” then Zuckerberg can’t appeal to Linus’s law any longer.

2. For “the truly catastrophic science fiction scenarios for humanity”, I think the appeal to Linus’s law mistakes the nature of these concerns. To my knowledge, many such scenarios postulate that the risks come from the unexpected emergence (or revelation) of radical new model capabilities at some stage of training, coupled with inadequate ‘alignment’ and ‘evals’, leading to sudden loss of control of an AI system somewhere in the ecosystem, with catastrophic consequences. By the time such an event occurs, it’s too late to file a bug report.

Researchers worried about catastrophic risks from AI seem to think that increasing access to advanced AI straightforwardly increases risks since it increases the number of opportunities for some developer or organisation somewhere in the ecosystem to recklessly push the capabilities of a model and unintentionally lose control of the system leading to this kind of harm. To put it another way, “given enough fingers, all catastrophes are shallow.”

Intentional harms. Zuckerberg further breaks down intentional harms by the amount of resources accessible to the bad actors, compared to the resources accessible to people whose job it is to protect against those harms.

Let’s start with the argument for ‘Open Source AI’ promoting safety against small-scale bad actors.

“

I think it will be better to live in a world where AI is widely deployed so that larger actors can check the power of smaller bad actors.

I can't make any sense of this as an argument. Say what you will about the closed AI world (again, I am not a fan of OpenAI), but the most likely reason a small-scale bad actor got a dangerous AI system is because a large-scale actor built one and released the weights. Zuckerberg's plan simply worsens this problem and then claims to solve it with the logic of [“the only way to stop the bad guy with an AI is with a good guy with an AI”](#).

It is also possible for small-scale bad actors to misuse the restricted access they have to AI systems offered by closed providers. For example, the safeguards enforced on OpenAI's models can also be bypassed, and OpenAI offers its own fine-tuning service. OpenAI presumably monitors usage of their systems for dangerous behaviour. This surveillance is problematic from a software freedom perspective, but *some* surveillance may be necessary from a safety perspective in a world with bad actors. Meta AI can't monitor usage in principle. Are we sure that is the right balance?

The threat model from state-level bad actors—the explicit example given is China—is slightly more coherent. Zuckerberg argues that the only way to stay ahead of such adversaries is to lean into decentralised and open innovation and retain a perpetual first-mover advantage. He claims these actors will be able to keep up with slower, closed AI development easily through sophisticated spy networks infiltrating the apparently hopeless security in Silicon Valley.

I don't really have the background to critique this with confidence, so I'll merely point out my uncertainties about it: Why couldn't Silicon Valley's security be improved if the world order was at stake? And how is it possible that on the one hand, all of the benefits of open source will apply to developers in democratic nations, but somehow developers working with adversarial state-level actors won't be able to keep up with the technology?

Relative vs. absolute safety standards. The heading of the section is 'Why Open Source AI Is Good for the World', but Zuckerberg only puts forward the following case:

“There is an ongoing debate about the safety of open source AI models, and my view is that open source AI will be safer than the alternatives.

In other words, Zuckerberg merely argues that his 'Open Source AI' is *better for the world than closed-source AI*. As I have outlined, I don't buy Zuckerberg's argument for *relative* safety. But what is worse is that he actually needed to make an even harder case, for *absolute* safety, to show that the technology is 'Good for the World.'

Fast-forward to a future where Zuckerberg's 'Open Source AI' is ascendant. It powers as many services as Linux servers and as many user devices as Android. The world is realising the vast potential benefits of the new technology. The world is also realising its harms. Is this world better off than our world today? Clearly, that entirely depends on how serious the harms turn out to be. Were they avoidable, or were they catastrophic? As far as I can tell, nobody knows how those harms are going to play out.

An alternative way of reading this letter is a scathing critique of Facebook and Instagram—these being examples of Meta products that promise benefits to society while working hard to trap them into a closed ecosystem and using closed-source AI to subtly mine their behaviour and undermine their psychology over time, fuelling an advertising machine that is the very reason “openly releasing Llama doesn't undercut our revenue, sustainability, or ability to invest in research like it does for closed providers.” If there's one topic on which Zuckerberg has no credibility, it's whether some technology is a net good for society.

§ The path forward is not so clear

One way of misreading my critique is to read the parts about software freedom and conclude that the Llama 3.1 release is *not open enough*. Wrong! Depending on how hypothetical AI risks play out, more openness could spell our doom.

Another way of misreading me is to ignore the fundamental importance of software freedom and conclude that OpenAI's 'closed' approach is justified. Wrong again! Sam Altman, or Silicon Valley more broadly, cannot be trusted with such a transformative technology. We can't settle for the level of centralisation that defined the Facebook era going forward.


If you think there's a contradiction here, you have understood my inner turmoil. I grew up, academically speaking, fascinated by Richard Stallman and enraged at the artificial barriers placed on technology all around me, not to mention the destruction left in the wake of the networks and platforms of the past decades. I am also a self-described AI safety researcher, building an academic career based on the premise that catastrophic risks from future advanced AI are more than science fiction.

How do I reconcile these two realities? The resolution is that yes, both of these problems could be real, and if so, that just means that *the path forward is unclear*. Maybe we need to slow down and spend time carefully thinking about and testing different approaches. Maybe we need to find out what we can do to advance the technology *selectively* in the direction of robustifying society, [as Vitalik Buterin proposes](#). Maybe this is not a problem that technologists should be the ones trusted to solve.

I don't have the answer yet. All I know is: Zuckerberg is wrong to confidently assert that "Open Source AI Is the Path Forward."

Move fast, break everything,

MFR

-
- 
1. Thanks to Daniel Murfet and Adam Dorr for helping me collect these thoughts at [a seminar in the open metaverse](#), and afterwards. Thanks also to Usman Anwar, Billy Snickers, Athir Saleem, and Jesse Duffield for helpful discussion.↵
 2. It's not stated in the letter, but it's a sound business strategy for Meta to promote competition in AI models because [turning AI into a commodity allows Meta to dominate supply chains](#) in AI-powered social media, not to mention the metaverse.↵
 3. I can *maybe* see a case for calling the weights of the model the 'source'. If you follow Andrej Karpathy's analogy that ML models are '[software 2.0](#)', where the learned weights correspond to the source code of traditional software, and they are written by learning algorithms rather than by traditional human developers, then the complaint that a developer can't effectively modify a system by directly modifying weights represents a failure to complete the translation. The developer should use a learning algorithm to modify the weights—as in distillation or fine-tuning. This is exactly what Zuckerberg seems to be proposing. Even though I can see this rationale, I refuse to use the term 'open source' in this way and in this post I try to make the case that whatever you want to call it, Llama 3.1 is not open enough to achieve Zuckerberg's vision.↵
 4. A counterpoint: My colleague Daniel Murfet points out that the Llama 3.1 report reveals the crucial role played by the previous generation of LLMs in the preparation of training and tuning data used to train the 3.1 herd. One could extrapolate this pattern to future generations and claim that by releasing the weights of Llama 3.1, Meta has released a fundamental tool necessary for developing the next generation to the open source community. My argument stands because while this may be a necessary component, it is not alone sufficient.↵

built with markdown, pandoc, make, and matte.css
last compiled: 2024-07-31

[top](#) | [home](#)