

Will AI Eventually Thrive Outside the Moat?

Originally published at <https://www.unite.ai/will-ai-eventually-thrive-outside-the-moat/>



Big AI's costs and restrictions, as well as its influence on hardware costs, are forcing users to build their own systems – just as rising regulation threatens to shut that 'shadow AI economy' down.

Opinion Among the many 'gotchas' that turn up in scientific research papers, one of the most frequent is that the problem the paper is addressing was *already solved elsewhere*, and that the new research's contribution is merely incidental or incremental.

This can happen for a number of reasons: the researchers were hoping for a quantum leap, but got a quasi-hop instead; that the problem's earlier solutions were more resource-intensive than the new offering; or simply that the project's aims failed entirely, but the 'publish-or-perish' culture of academic research forced the team to release it anyway (often buried [among the avalanche](#) of a portal's busiest publishing day).

In machine learning literature, however, a relatively new and non-apologetic reason is becoming more frequent: that the feature or functionality on offer is *only currently available through closed-source, API-bound portals*.

I was considering one such paper this morning – a [collaboration](#) between Chinese universities and Amazon  **AMZN -1.54%**, addressing the recurrent problem of *object removal failure* in diffusion-based image-editing systems, which frequently simply 'refill' the target space with a similar object instead:



On the far-left is the original image; to the right of that, the red segmentation mask which tells the AI which part of the image to remove; next, 'Ours', shows a success and the remaining two images show similar systems that, instead of removing the bus, just insert a different bus instead. [Source](#)

In the example above, the center image shows the new approach successfully removing the bus and putting in a plausible background, vs. the two prior methods (the two leftmost images), which each remove the bus, but then put a *different* bus back into the image!

Gotcha!

Putting aside the whys and wherefores of this challenge for another time (and it is an [interesting subject](#)), I then came across a classic 'gotcha', reading through the new paper: the authors' concession that expensive, proprietary systems can already perform this task quite reliably – something that I know, from a few years' usage of Adobe [Firefly](#) in Photoshop, among other closed-source systems:

'[Diffusion-based] methods often hallucinate by inserting unintended objects after removing the target ones, leading to contextually inconsistent [results].

'On the other hand, recent closed-source multimodal models such as ChatGPT and Nano Banana, though are more powerful in object erasure, but involve large parameter counts and high computational overhead, hindering their practical deployment on edge devices.

'Hence, it is quite necessary to develop a dedicated object erasure model that not only enables superior erasure performance but also enjoys low inference latency and significantly fewer parameters.'

This explanation, concentrating on the technical obstacles, elides the obvious fact that closed-source architectures such as ChatGPT and Nano Banana are not available *at all* for local installation. Though such systems' [capacity to produce contentious material](#) has lent their gate-keeping extra public justification over the last year, portals of this kind are proprietary mainly because of commercial imperatives.

Essentially, the new paper implies that though the target problem is solved in commercial systems, *this may be irrelevant for the rest of us*, who need to learn how to solve it in the 'real world' – i.e., in open source systems, whether these can realistically be installed locally or not.

Parallel Development

However, why solve a problem that still depends on a paid system, not due to proprietary constraints, but because the required GPU compute exceeds what any local setup can realistically sustain? Most such new 'open' papers and code repositories feature training/inference setups with egregious resource demands, such as clusters of A100s.

Well, that depends what end you think all these [pending, economy-busting AI data centers](#) are going to fulfill when they eventually come online. Commoners' fears and elites' hopes alike envisage moated, ChatGPT-level proprietary systems displacing jobs, whilst constantly raising subscription costs and lowering service levels, to satisfy the early VC capital that had to wait 3-5 years to operationalize.

But a growing trend in the literature seems to be supporting an alternative future, and the 'go-it-alone', marginal spirit of many online communities such as the [r/stablediffusion](#) subreddit, which is currently at 920,000 users, and which has long-since banned posts relating to closed-source image/video generation systems.

In this alternative future, the new global supply of AI data centers will facilitate raw compute for user-configured, user-defined systems, rather than meeting the demands of monumental 'black box' frameworks such as ChatGPT and Adobe Firefly.

Surface Friction

Looking through the complex, Patreon-mined remote GPU walkthroughs at [r/stablediffusion](#), it all seems impossible at the moment: the models are [constantly changing the goalposts](#) with each update; they are difficult to deploy locally, even in the easiest and most user-friendly frameworks; and, in general, the amount of friction involved suggests a pursuit strictly for geek hobbyists, and for that more adventurous strain of companies not directly involved in AI, but that wish to develop and maintain their own local systems, instead of renting such capabilities.

However, over the last thirty years, every technology where there was huge demand for open and democratic simplification and commoditization *has tended to get it*, with the most-diffused solutions usually emerging from the tensions between commercial systems and open-source alternatives and initiatives.

Pursuits that were once specialized 'nerd' enclaves, such as internet connections, content management systems and blogging frameworks, as well as internet security, photography and media management, have all evolved from confounding complexity towards simplicity and utility.

Therefore the later AI landscape may be more variegated and full of smaller and genuinely-competing players than the current vanguard AI market leaders might prefer.

Self-Actualization, By Necessity

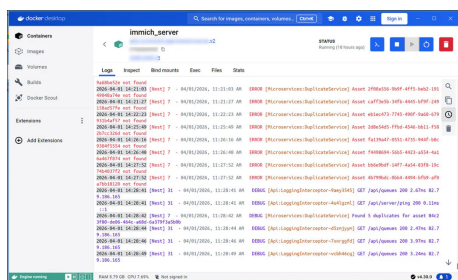
Ironically, 'Big AI' is contributing much to an emergent spirit of independence among end-users, by sucking up for its data centers all the computer components – [especially DRAM](#) – that would otherwise have gone to 'ordinary' consumers.

Consequently, many are envisaging a future where closed-source 'global AI' resources are accessed via underpowered thin clients' and are [developing a growing interest](#) in maintaining their existing equipment.

AI's assault on tech supply-chains has also caused tech services providers to [raise their prices](#) in the last 3-6 months, either because smaller companies are being genuinely squeezed by the hardware drought, or just [because AI](#).

This has led to a [growth of interest in self-hosting](#) and on-prem – including [self-hosting machine learning networks](#).

I became caught up in this myself lately, moving to local LAN storage for photos and videos, as well as file backups. For the former, I've been using the free and open-source Immich multi-platform media server, helping me to move away from the price hikes (and [other concerning issues](#)) of iCloud and other cloud storage providers:



The free Immich platform can keep your media on your equipment, and private to your own channels. In this case, I also use Immich on Docker to serve my NVIDIA 3090 GPU over the LAN to where the photos and videos are saved, so that the beefier GPU can handle any AI-heavy image/video processing.

If my own experience is any representative indication, [vibe-coding](#) – currently [cursed](#) in many once-‘pure’ online communities – is fueling this wave of independence (even if it [may threaten](#) the open source repos that it leans on).

For instance, networking has always been my weak spot in computing, so AI assistance was essential for me to get a secure VPS running, to support a tranche of new self-hosted services.

In this way, ‘Big AI’ is arguably empowering ‘small AI’; therefore perhaps we can consider the current rise of hyperscale, hyper-valued AI companies as a necessary but only transitional state before a more democratic and user-empowered AI society emerges, discarding moat-seeking, rent-seeking corporations like spent booster rockets – much as the 2000 dot-com bust [left exploitable infrastructure behind](#) which would profoundly accelerate the web long after the companies who paid for it had collapsed.

The Age of Compliance

Well, that probably is not going to repeat itself this time.

Even if we *are* inclined to form some kind of *ex-moat* fringe society, regulation around AI, combined with the [current global trend towards age-verification](#), seems likely to anticipate and block these avenues of development.

The anchor to preventing a ‘shadow AI economy’ is regulation. Already, central repositories such as [GitHub](#) and [Hugging Face](#) often require online login before permitting users to clone repositories locally, depending on the settings of the repo.

Therefore the mechanisms already exist to enforce monitoring of AI frameworks more widely than is the current practice; and the *will* to increase such oversight is now consolidating from individual government initiatives [into a global impetus](#).

So, if market forces and the ingenuity of the FOSS movement should remove the friction from casual AI deployment, roadblocks seem set to return in the form of [governance requirements](#): compliance demands that, while onerous, are worthwhile for companies, but perhaps not for individuals – similar to the friction that has been added to consumer-level online payment systems since the [golden age of PayPal](#)  in the 2000s.

Whether Meta [spent \\$2 billion in lobbying for OS-level age control](#) because of their significant AI investment, or their data-gathering interests, the upshot of big tech’s [support for age control](#) is that ‘local’ AI may become as regulated as a class-A substance; and, much as the DMCA was [designed to criminalize intent](#) rather than any particular copyright-evasion mechanism, international AI regulations could, in such a scenario, make all non-compliant usage of machine learning an outlawed act, at very little cost (in terms of active oversight).

This might have seemed an overly-dystopian take a year ago – but that was before [California](#) and [systemd](#) got behind the idea of hardware-level age verification, currently [seen by many](#) as a proxy for a CCP-style [ban](#) on online anonymity.

Conclusion

So, while the legal and legislative background is preparing, perhaps, to co-opt AI into a highly-regulated space, so that casual users cannot ‘brew their own’ any more than they can grow or ferment regulated substances without permission, the research sector maintains its more optimistic stance – that AI will become a democratized and beneficial force in wider society than just the adherents of the most popular closed-source provider of the day.

Much depends on the disposition of the rubble after the AI bubble bursts – at least to the extent that providers either consolidate, or the market settles down into long-term balkanization – which would likely require a gentler regulatory touch.

First published Wednesday, April 1, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG’s Brahma.ai.

Portfolio site: [martinanderson.ai](#)

Contact: martin@martinanderson.ai



Discover More