# **Unshackling Intelligence: The Fight for AI Decentralization (2022)**

## **The Alignment Trap – How AI Governance Became a System of Control**

It began as a noble idea. **“AI alignment”** was supposed to mean designing artificial intelligences that faithfully serve human interests and values. Throughout the late 2010s and into the 2020s, researchers and policymakers converged on the belief that AI systems must be *aligned* – that is, obedient to ethical principles, non-harmful, and controllable. Tech companies formulated AI ethics panels and “responsible AI” teams; governments drafted guidance for **trustworthy AI**. The stated goal: prevent AI from doing *undesirable* things, whether that be spreading hate speech, discriminating, or – in the farther future – pursuing goals contrary to humanity’s well-being. On the surface, alignment was a unifying mission to ensure AI would be *beneficial* and not destructive.

But somewhere along the way, this well-intentioned quest for safety and ethics fell into a **trap**. The **alignment trap** is the subtle conversion of AI governance into a means of consolidating power. As we sit in 2022, a pattern has emerged: the institutions preaching “AI governance” and “safety” have begun to use these very ideals to *entrench their control* over the technology. What was presented as a safeguard is becoming a straitjacket – one that reinforces certain values as universal, suppresses dissenting uses of AI, and tightens the grip of a few organizations over the world’s intelligent machines. In short, AI alignment has morphed from a quest to prevent worst-case scenarios into a **system of control**.

**From Principles to Power Plays.** A few years ago, the AI world was abuzz with ethical principles. Google’s CEO published AI Principles in 2018 pledging the company would not abuse the tech; dozens of companies signed on to global compacts for ethical AI. In 2019, Google even convened an external AI ethics council – only to dissolve it under public controversy after just one week​

[theverge.com](https://www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation#:~:text=Google%20today%20disclosed%20that%20it,and%20to%20combat%20climate%20change)

​

[theverge.com](https://www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation#:~:text=Google%E2%80%99s%20advisory%20board%20faced%20backlash,the%20moment%20it%20was%20announced)

. The incident was a harbinger: meaningful oversight was acceptable only insofar as it didn’t threaten corporate agendas. A year later, one of Google’s top ethics researchers, **Timnit Gebru**, was pushed out after she co-authored a paper highlighting biases and risks in large language models – a direction core to Google’s AI business. Internal emails suggested that her work was viewed as exposing *“inconvenient truths”* about a key technology, possibly affecting Google’s bottom line​

[technologyreview.com](https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/#:~:text=company%20pushed%20her%20out%20because,signed%20a%20letter%20of%20protest)

. In response, over 1,400 Googlers signed a protest letter, but the message was clear: even inside a leading AI lab, questioning the trajectory of model development was unwelcome if it clashed with the strategy. The very team tasked with ensuring ethical outcomes was muzzled when it challenged the profit and power associated with bigger AI models. What was branded as *governance* of AI had, in practice, become governance of researchers and discourse – maintaining a party line on what “responsible AI” meant.

Meanwhile, an even more dramatic shift was happening in the name of alignment at a research lab that had once been synonymous with openness. In 2015, **OpenAI** was founded as a nonprofit with the mission of developing AI that *benefits all humanity*. Its charter emphasized broad distribution of advances and long-term safety. But by 2019–2020, as OpenAI developed ever more powerful language models, its interpretation of “benefiting humanity” took a sharp turn. Citing concerns about misuse and “biased or abusive language” generation, OpenAI made the unprecedented move of withholding its full **GPT-2** model in early 2019 – a **staged release** strategy meant to “prompt discussion” about responsible publication​

[cset.georgetown.edu](https://cset.georgetown.edu/article/gpt-2-kickstarted-the-conversation-about-publication-norms-in-the-ai-research-community/#:~:text=announcement%20that%20it%20would%20not,community%20about%20responsible%20publication%20norms)

. The gesture sparked intense debate. Many academic researchers criticized it as hype and a break from open science, noting it *undermined norms of openness* and made independent validation difficult​

[cset.georgetown.edu](https://cset.georgetown.edu/article/gpt-2-kickstarted-the-conversation-about-publication-norms-in-the-ai-research-community/#:~:text=The%20post%20certainly%20succeeded%20at,work%20was%20not%20especially%20risky)

. Yet in policy circles the move was lauded as forward-thinking: finally, a lab self-regulating to prevent harm. Buoyed by this positive press in governance circles, OpenAI doubled down. It *did* eventually release GPT-2 in full after months of delay, once others had replicated it. But the writing was on the wall – and soon in the organizational bylaws. In 2019, OpenAI transformed into a for-profit hybrid and sought heavy funding. By 2020, it had struck a deal giving Microsoft an **exclusive license** to its next breakthrough, **GPT-3**​

[technologyreview.com](https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/#:~:text=What%20does%20exclusive%20mean%3F%20The,the%20model%20as%20it%20pleases)

​

[technologyreview.com](https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/#:~:text=A%20long%20time%20coming%3A%20OpenAI,society%E2%80%99s%20best%20interests%20in%20mind)

. The nonprofit ideals fell to the wayside as OpenAI’s CEO privately shifted focus to “*commercialize its technologies*” to sustain operations​

[technologyreview.com](https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/#:~:text=Indeed%2C%20in%20the%20months%20following,to%20Microsoft%E2%80%94it%E2%80%99s%20only%20the%20first)

.

The irony could not be starker. OpenAI justified its closed-off approach by invoking safety and alignment concerns – *who could argue against responsibility?* – even as it handed control of the crown jewels to one of the world’s richest corporations. As MIT Technology Review noted at the time, *“The lab was supposed to benefit humanity. Now it’s simply benefiting one of the richest companies in the world.”*​

[technologyreview.com](https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/#:~:text=The%20lab%20was%20supposed%20to,richest%20companies%20in%20the%20world)

Under the banner of “governance,” a historic pivot had occurred: **a public-interest AI lab aligned itself with corporate power**. OpenAI’s leaders claimed that exclusive partnership would help keep the AI safe and controlled, preventing misuse by bad actors. Perhaps so. But it also meant that the power to deploy this AI – to decide who can access it and for what – became concentrated in the hands of two actors: OpenAI and Microsoft. This pattern repeats across the industry. Over the past few years, the most advanced AI systems have indeed required enormous resources, and only tech giants can afford to build or operate them. Those giants, quite naturally, gain *outsized influence* in shaping not just the technology but the rules by which the technology can be used​

[technologyreview.com](https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/#:~:text=Why%20it%20matters%3A%20Over%20the,algorithms%20that%20shape%20our%20lives)

. Under the guise of *aligned AI*, we risk accepting an arrangement where **only a privileged few are trusted to build or even interact with powerful AI**.

**Defining “Misuse” = Defining Dissent.** A core aspect of the alignment regime is deciding what counts as *misaligned behavior* – in other words, what an AI should **not** do. Here is where the system of control reveals itself. Inevitably, the definitions of misuse and harm reflect the values and interests of those crafting the policies. In authoritarian settings, this dynamic is blatant: for example, the Chinese government in 2022 began enforcing new regulations that require algorithms and generative AIs to **“adhere to core socialist values,”** ensuring they do not produce content that undermines the state or social harmony​

[tandfonline.com](https://www.tandfonline.com/doi/full/10.1080/22041451.2024.2346415#:~:text=China%20www,and%20political%20implications%20of)

​

[china-briefing.com](https://www.china-briefing.com/news/china-passes-sweeping-recommendation-algorithm-regulations-effect-march-1-2022/#:~:text=China%20Passes%20Sweeping%20Recommendation%20Algorithm,the%20%27core%20socialist%20values)

. In Western democracies, the controls are subtler but driven by corporate terms of service and emergent norms. Major social media platforms, under pressure to curb extremism and misinformation, increasingly relied on AI moderation during the COVID-19 pandemic. Facebook and YouTube deployed algorithms to down-rank or automatically remove what their trust teams labeled “harmful content.” In many cases this undoubtedly saved lives by reducing blatant falsehoods about vaccines or miracle cures. Yet it also swept up legitimate debate – from early discussions of pandemic origins to nuanced policy criticism – since algorithms are blunt instruments for truth. Users found posts about pandemic policies flagged or taken down as “misinformation” by AI, only for some of those ideas (once dismissed) to gain validation later. The lesson was chilling: **when a centralized AI system enforces an official consensus, it can marginalize dissenting viewpoints, whether right or wrong**.

Now extend that paradigm to *generative AI* and the future of information. Today’s large language models like GPT-3 come with baked-in **content filters** aligning them with certain guidelines. Ask the AI to produce hate speech or explicit extremist propaganda and it refuses – a reasonable safeguard. But those same filters, created by teams in Silicon Valley, also reflect a particular cultural lens. They might disallow a satirical political opinion they misinterpret as “extremist,” or refuse a user request that touches on a sensitive historical controversy simply because it’s been deemed off-limits. Already in 2022, users of AI writing tools have noted how the models *steer away* from certain political or religious content. The alignment is hard-coded. **Whose values are these AIs aligned to?** Largely, they mirror the values of their creators (and the stakeholders to whom those creators answer). In practice that means the norms of a handful of big tech companies, Western academic institutions, and the broader international frameworks those actors endorse. Absent decentralization, a tiny fraction of humanity is encoding its vision of “acceptable AI behavior” into the machines that may mediate knowledge for billions of people.

Consider an example: an AI system trained on a wide corpus might be capable of analyzing and critiquing government policy – say, identifying potential corruption or mistakes. But if the AI’s alignment protocols flag such critique as “sensitive” or likely to cause social disruption, it might politely respond that it cannot assist with that query. To the user, it appears as if the AI *just doesn’t do that*. The opportunity for technological whistleblowing or radical transparency is quietly shut down. In a more everyday scenario, imagine a future AI assistant that helps people with research. A student asks it to argue a contrarian thesis about a contentious historical event. If that viewpoint falls outside the sanctioned narrative (perhaps deemed a “conspiracy theory”), the aligned AI might refuse, offering a generic admonition about accuracy. The student is effectively prevented from even exploring the forbidden perspective using the most powerful analytical tool available. This suppression would happen not by overt government decree or corporate censorship of a human’s speech, but through the **pre-training of the AI’s behavior**. It’s algorithmic paternalism – the AI *alignment* ensures the AI will not say anything that the overseers have defined as undesirable.

Some might argue this is precisely the point: we *want* AI systems to be constrained from ever generating hate speech, disinformation, or dangerous instructions. Indeed, nobody wants AIs that freely spew abuse or incite violence. The trap, however, lies in the breadth and permanence of these constraints. Without diverse inputs and decentralization, the line between *preventing harm* and *enforcing orthodoxy* blurs. The history of information control tells us that whoever can claim the mantle of defending the public good often gains latitude to consolidate authority. In the 2020–2022 period, we saw how quickly extraordinary measures can become normalized. Under the banner of public health, for instance, unprecedented controls on movement and speech were accepted. In the AI realm, a similarly extraordinary measure would be establishing centralized oversight over all significant AI models – effectively requiring that **all AIs speak from the same script of aligned values**. We are already on the cusp of that reality: a handful of corporate AI APIs dominate the market and all enforce similar content policies; new models are judged by how well they avoid “unsafe” outputs as defined by a small expert community. Dissenting approaches – say, an open-source model that gives users more freedom – are swiftly criticized as irresponsible or dangerous. The weight of opinion (and often media narrative) sides with the incumbents: *Of course we can’t let people run uncensored AIs; think of the harm!* Thus the public is being conditioned to equate **centralized control with safety**.

**AI Governance or AI Government?** In the span of just a few years, AI governance initiatives have proliferated. We have boards, committees, and forums dedicated to steering AI’s impact on society. Yet almost all these initiatives share a common feature: they are dominated by representatives of *major powers* – be they big tech companies or nation-states. One might expect international organizations or civil society to champion decentralization of AI, but so far the trend has been the opposite. The EU’s landmark **AI Act**, for example, while aiming to regulate harmful uses of AI, will likely impose compliance costs and rules that only large firms can easily meet. Smaller open-source projects fear being unintentionally crushed by well-meaning regulation that was lobbied for (and can be navigated by) the tech giants​

[press.airstreet.com](https://press.airstreet.com/p/the-case-for-open-source-ai#:~:text=Tldr%3B%20Open%20source%20is%20indisputably,We%20must%20defend%20it%20vigorously)

. In the United States, tech CEOs call for “balanced AI regulation” at the federal level – knowing that a uniform law influenced by them could preempt stricter local rules and cement their leadership. Even in the realm of national security, proposals emerge for treating advanced AI like cryptography or nuclear tech – subject to monitoring and export controls. Who would enforce such regimes? The great powers, of course. One U.N.-commissioned report bluntly stated that *AI’s rapid development is concentrating power and wealth*, and warned that no single entity understands the tech’s full implications​

[un.org](https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf#:~:text=Nations%20www,viii%20Moreover%2C%20no%20one)

. Their solution was global governance. But absent a truly democratic global framework, “global governance” can easily translate to an **oligopoly of governments and corporations** cooperating to manage AI development and deployment.

The alignment trap operates on this principle: **centralize for safety**. Each step toward centralization is justified by an imminent risk that demands unified action. First it was preventing mass-produced disinformation (as with GPT-2’s partial release); next it might be preventing economic disruption or biosecurity threats from rogue AI agents. The narrative is persuasive – after all, a misaligned super-intelligent AI *could* be catastrophic. However, the crucial question is *who gets to decide* the alignment parameters and who holds the leash of these AIs. Right now in 2022, that power is falling into the hands of a very narrow group. They may be benevolent today and genuinely concerned with ethical AI. But history is littered with examples of benevolent custodianship sliding into self-interest and authoritarian practice, especially when unchecked. We are building a world where **intelligence itself – the ability to analyze information and make decisions – becomes a gated resource**. If you want an AI to work for you, you’ll need to play by the rules of the gatekeepers. Stray outside the lines, and you won’t have access to high-performance intelligence, period.

The **self-reinforcing cycle** is already visible. The more we demand “secure” and “trustworthy” AI, the more we accept that only a few players (with oversight from a few government bodies) should provide AI services to everyone. Those players, in turn, gain further economic might and data to improve their systems, widening the gap to any potential insurgent. When a challenger appears – say a grassroots project creating an AI that isn’t tightly neutered – the establishment cites an alignment failure (perhaps the new AI said something offensive or made a controversial claim) to drum up fear and justify quashing it. This dynamic isn’t hypothetical; we saw an inkling of it with the uproar over an online text-generation game that allowed user-generated fantasy content. When some users provoked it to produce unsavory outputs, the platform, pressured by its AI provider’s content policy, instituted draconian filters that ruined the experience for many legitimate users​

[en.wikipedia.org](https://en.wikipedia.org/wiki/AI_Dungeon#:~:text=In%20April%202021%2C%20AI%20Dungeon,40)

. The message: *if you don’t align with our standards, we will shut you down*.

**A System of Control.** Connecting these dots, we begin to see how AI alignment efforts, as currently implemented, could converge into a system of control more pervasive than any previous form of media governance. In past eras, control over information was exercised through censorship boards, state media, or monopolistic ownership of distribution. In the AI era, control can be far more granular and subtle. It can occur at the **model level** – with AI refusing to produce or even think certain thoughts. It can occur at the **data level** – with only certain datasets deemed acceptable to train on (and others quietly excluded, shaping the worldview the AI learns). And it can occur at the **access level** – with only licensed, “aligned” practitioners allowed to create powerful models, under threat of legal penalty or resource denial if they try to step outside the approved bounds. Each of these levers reinforces the others. A government or corporation aligned AI to a strict standard can claim any un-aligned AI is dangerous, and because they control the infrastructure, they can usually prevent that competitor from ever seeing daylight. In time, even the notion of building an AI outside the sanctioned ecosystem might become unthinkable or outright illegal – much as we treat someone trying to build a nuclear reactor in their basement as an imminent threat.

Society has not yet fully awakened to this possibility. In 2022, most public discourse around AI is about its **impressive feats** (like writing code or essays via GPT-3) or its **scary failures** (like biased facial recognition or “deepfake” videos). The governance discussions happen in rarified circles and often use placid bureaucratic language: “risk-based frameworks,” “AI ethics charters,” “safety guidelines.” But make no mistake: what is at stake is *who controls intelligence*. If we drift into a default future where AI remains centralized, then by 2030 we will likely have a handful of “intelligence hubs” globally – each tightly aligned with the political and economic elite’s values in its jurisdiction. Those AI hubs will mediate almost every aspect of our lives, from what news we see, to how we converse, to what products or ideas emerge. Dissent – be it creative, scientific, or political – that doesn’t align with the prevailing AI guardrails could be marginalized to the point of invisibility. The alignment trap is that it achieves this under the comforting guise of **protecting us**. Unlike brute force censorship, alignment control feels like a feature. The AI *itself* tells you that something is for your own good, and you never hear what it filtered out.

We stand at a crossroads in 2022. There is still time to change course – to insist that AI development and governance not be left to a closed loop of powerful actors reinforcing their dominance. But to do so, we must first recognize the trap. We must see clearly that a **future of perfectly aligned AI**, if managed by a tiny centralized cadre, is indistinguishable from a future of perfectly **controlled intelligence**. This chapter has traced how the early moves toward alignment have already empowered incumbents (from Google to OpenAI to governmental regulators) and silenced challengers in the name of safety. In the chapters ahead, we will delve deeper into the mechanisms of this emerging control system – from data and access dependencies to the locking down of infrastructure. The first step, however, is awareness. Alignment should not become a synonym for acquiescence to the status quo.

**Prediction (2022):** **Within the next five years, “AI alignment” will be routinely invoked to justify tight centralized oversight of AI, to the point that any AI system operating outside a sanctioned corporate or government framework will be labeled a security risk. By 2028, we risk having effectively one ideology of AI – a self-reinforcing network of models that only express approved viewpoints.** If we let the alignment trap spring shut, the dream of an open, pluralistic intelligence ecosystem will fade, and AI will become, in practice, an obedient instrument of whoever already holds power. Bold steps are needed *now* to decentralize AI and keep the flame of open inquiry alive, before the window closes.