*Specific feedback I’m looking for*

* *Writing style: I'm unsure whether to go more for a entertaining/funny style or be academic in tone.*
* *Any improvements to writing style? Esp high-level observations about type of writing/formulations you like*
* *Are there any other vectors you can think of*
* *Any important considerations I forgot*
* *Anything I could add to make the post more valuable*
* *Feedback on the Introduction is prob extra valuable*
* *Anything else you do/don’t like*

16 Vectors for influencing Frontier AI Companies

Frontier AI Systems are being built at large AI Companies. So if you want to make sure the AIs are safely developed and deployed, you need to influence these companies. At some point in your Theory of Change you need to say “and thus OpenAI did a better job”. In this post I will list all the ways in which AI Companies can be influenced. If you can’t point to one of them as part of your pathway to impact you’re either achieving nothing or (more likely) should correct me in the comments.

I’ll go through different vectors for influencing AI companies. Thereby, I will only include *final* nodes that directly influence the company without being mediated by other nodes (eg raising public awareness doesn’t count since it only indirectly influences companies through other mechanisms). I’ll add examples, some spicy takes on their impact potential and who is suited to pull on these levers.

Exception: There are legitimate interventions for reducing risks from AI that don’t go through AI companies. They are targeted at making society more resilient to the transformation caused by AI or the risks it poses. Examples would be work on [Gradual Disempowerment](https://gradual-disempowerment.ai/), [d/acc](https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html#dacc) or building [Tools for Existential Security](https://www.forethought.org/research/ai-tools-for-existential-security). Furthermore, I’m assuming that AI Companies are the leading actors developing AGI instead of a [CERN for AI](https://cfg.eu/building-cern-for-ai/)-type of structure.

Limitations: This analysis doesn’t go in depth on any vector and doesn’t analyse their effectiveness. *I’ve thought about this for 1 day. I’m confident each of the vectors is a real lever for influence (although they differ in effectiveness), but might have missed some. Esp analyses in the “Regulation” section likely have important gaps, as it is not my expertise*

# Direct Work

## Join them

Working at the AI companies means you can be in the room where decisions happen. You can play internal politics, make new projects happen or just be in a decision-making position. This gives you sway over crucial decisions about compute allocation, release strategies or research priorities. You can also contribute by implementing solutions and carrying out projects that have a positive impact, for example by being an individual contributor in a safety team. Lastly, insiders have access to sensitive information that they can leak to governments or the public if necessary. These positions are clearly very impactful as they are close to the action, but also bring risks like accelerating AI development or being corrupted by power/money/social influence (more [discussion](https://80000hours.org/career-reviews/working-at-an-ai-lab/)).

But not everybody can (or should) work in an AI Company. As of 2025 Deepmind, Anthropic and OpenAI have a combined 6600 highly-competitive positions. Furthermore, it is important that some people play outside baseball to put pressure on the companies and give safety-conscious employees internal leverage. So what should the rest of us do?

*Who?* Potential employees

## Build something for them

AI companies are lazy and have limited capacity, so for a lot of important things they won’t have time for it. You can help them by building a tool they can use, inventing a method they can implement or curating a dataset. In this way you can differentially accelerate good developments of AI.

* Build open source tools or paid products the companies can use. For example you could build a tool for monitoring agent actions that companies can easily deploy.
* Invent new methods that can be implemented inside companies. For example more robust safeguards or alignment algorithms.
* Build datasets they can use for training or evaluation. AI companies are hungry for data and thus likely to use your curated, high-quality dataset. So this can be a great way to differentially accelerate or measure capabilities you care about.

This lever is open to anyone with the ability to build the appropriate technology. Note, that your goal is usage by the companies, so you should also spend significant effort to make sure they are aware and potentially implementing it.

*Who?* Researchers, Entrepreneurs, Engineers

## Collaborate with them

This is essentially a more direct version of building something they can use. It’s preferable in that you have a higher chance that the company actually uses what you built and that you get some inside information about where you can meaningfully help. It’s undesirable as you might face a balance between honest communication and good relationship with the company or might be used for safety washing. Examples of collaborations are external safety testers (like METR) or data collectors (like ScaleAI). Collaborating with researchers inside AI companies also gets you the mentioned benefits.

*Who?* Researchers, Entrepreneurs, Engineers

# Regulation

## Regulate them

Regulation is the most salient strategy for shaping the behavior of AI companies from the outside. A lot of strategies aim to have their impact through increasing the likelihood and effectiveness of AI regulation being passed. Politicians, staffers, think tankers, civil servants, advocates and some technical research all aim to influence AI Companies by getting governments to create, change or remove legislation that determines how AI Companies need to or are incentivised to act. For example the EU AI Act’s Code of Practice is already demanding a range of risk management steps AI companies need to follow. Notable legislative efforts (e.g. SB 53) are underway and could pass in multiple US states.

The country in which the AI company is based clearly has the most leverage for guiding action. But other governments also have leverage if their country is a large market (e.g. the EU), houses important suppliers (e.g. ASML in the Netherlands) or offers huge funding (e.g. UAE).

*Who?* Policy makers, think tanks, advocates, voters

## Take them over

Instead of slow, boring regulation, how about the government directly taking over control of AI Companies? While this seems outlandish now, it might become plausible in a world with rapid AI developments, large societal changes and obvious catastrophic risks. In such a world the US Government could well decide that it wants to steer this development instead of leaving it in the hands of tech leaders without democratic legitimacy. There are precedents of such as the US government taking control over private railroads or a failed seizure of steel mills during wartime. There might be legal backing in the Defense Production Act or after a National Emergency has been declared.

*Who?* Government Executive

## Sue them

Aside from government action, outside actors can also help to enforce regulations through litigation. The most prominent example is the New York Times suing OpenAI over copyright law, which will likely lead AI Companies to change their ways of accruing data. There are many plausible angles for litigations from privacy violations to product liabilities. Aside from “winning” in court, such litigation can also force transparency over internal documents (see [Musk vs OpenAI](https://www.lesswrong.com/posts/5jjk4CDnj9tA7ugxr/openai-email-archives-from-musk-v-altman-and-openai-blog) emails), stop or slow down releases and impose significant costs on the defendant.

*Who?* Anybody that has been harmed, Lawyers

## Set standards for them

The US has no comprehensive legislation on AI. In such a vacuum it is natural to look towards standard setting bodies and agreed upon industry best practices for guidance on how AI companies should act responsibly. Standards and best practices matter as they can turn into legislation, become procurement requirements or can limit liability. Furthermore, by following industry standards and best practices a company can signal that it is a responsible actor.

However, it’s important who sets the standards. While government led standards like [NIST’s AI Risk Management Framework](https://www.nist.gov/itl/ai-risk-management-framework) can set important directions, industry designed self-regulation can just be a way to prevent stronger government action. For example the [Frontier Model Forum](https://www.frontiermodelforum.org/) that facilitates cooperation between AI companies on finding best practices, information sharing and advancing AI Safety. However, it is financed and led by the AI companies themselves and thus unlikely to cause large changes in their behavior.

*Who?* Competitors, Standard Setting Bodies

# Improve Decisions

Employees in AI companies need to constantly make high-stakes and uncertain decisions. Which research direction to pursue? How much compute to allocate to which project? What release strategy to use? We should help them in making better decisions.

## Convince them

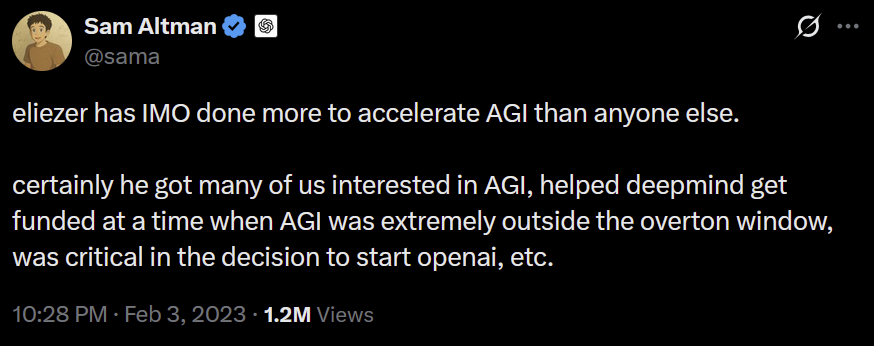
We can help people in companies by providing convincing arguments and evidence that improve the quality of decisions. For example researching weaknesses of a safety method, providing forecasts about AI development or introducing new strategic considerations.

Aside from producing new evidence and arguments, the right information also needs to reach the right people. That means there is an important role in filtering and delivering the right information to the relevant people inside of labs. Such a role could be filled by advisors, consultants or trusted internet sources (e.g. newsletters).

*Who?* Researchers, Advocates, Advisors, Anybody with access to employees

## Spread memes to them

One can attempt to seed and spread viral ideas that change one's way of thinking about the world. Through cultural dissemination these can reach researchers and decision makers at AI companies and influence their decisions and actions. Eliezer Yudkowsky, chief doomer and elite meme spreader, is a great example of this. His work was mostly impactful by popularizing the idea that superintelligent AI could spell doom for humanity.



*On the other hand, once memes are out they can have unintended consequences.*

*Who?* Anybody, respected figures, AI "influencers"

## Setting societal expectations for them

Workers at AI Companies are not rational robots, but social animals that are responsive to public opinion and moral judgements. Aside from being convinced by good evidence and arguments, they will also change their judgement based on the expectations and moral judgements of friends, family and the general public. If they are aware that everybody would hugely respect them if they did X and think they are monsters if they did Y, they are more likely to do X. Notably, this shouldn’t only look like guilt-tripping and criticism, but also include positive encouragement and admiration for doing the right thing.

*Who?* Anybody, Advocates

# Economic

## Invest in them

AI Companies depend on large investments to continue scaling. Actors who can provide large funding thus gain some leverage to influence the company as a condition of investment. For example, an investment by SoftBank into OpenAI was made conditional on the company changing its governance structure, which did cause OpenAI to attempt a (possibly illegal) for-profit conversion.

Investments can also be tied to voting rights or board seats. An early investment by OpenPhilanthropy in OpenAI enabled them to nominate a value-aligned board member (Helen Toner). While this didn't turn out great in hindsight (suspending Sam Altman for 4 days wasn't worth 30mio$), it looked for a while like an incredible bargain.

Unfortunately, as valuations of AI companies rise (OpenAI [reportedly selling at $500billion](https://www.cnbc.com/2025/08/15/openai-6-billion-stock-500-billion-valuation.html)), influence is restricted to actors that can move many billions in funding. Furthermore, "non-super-duper-rich" individuals cannot directly invest in AI developers as they are not publicly traded (OpenAI, Anthropic) or only a part of a much larger Tech Company (Google Deepmind, Meta AI).

*Who?* VCs, Investment Companies, very rich people

## Sell to them

AI Companies depend on a range of suppliers and providers of services. Suppliers can use this dependence as leverage to steer the company's actions. NVIDIA or TSMC hold especially large power as there are no competitors (90%+ market share), their products cannot be replaced by others and their products are essential to the operation of AI Companies.

Similarly, could compute providers (AWS, Oracle Cloud, …) hold large leverage, although they are more replaceable. For example Microsoft's deal to provide cloud compute to OpenAI bought them a (non-voting) board seat, access to all pre-AGI tech and likely granted them significant input in the development and release of GPT4.

While GPU/semi-conductor companies and compute providers have by far the largest leverage, there are also other suppliers:

* Data providers: Scale AI et al. label and collect high-quality data; Reddit or Publishing houses hold the rights to very useful training data
* Technical Infrastructure: MLOps platforms such as Weights & Biases or ways of disseminating AI models like HuggingFace.
* Integrations: Services that offer integration with the AI like Atlassian, Cloudflare or Paypal that offer [integration with Claude](https://www.anthropic.com/news/integrations)
* Other Infrastructure: Banking services, …

*Who?* See in bullet points above

## Buy from them

Companies need to earn money and will thus be responsive to customer demands. This aspect is strengthened by the similarity between products of different providers, which makes the cost of switching very low.

Single Customers can set financial incentives for AI companies by voting with their wallets or boycotting products. While these are popular tactics in the animal rights movement or environmentalism, they have much less use in AI Safety. However, for companies focused on B2C, large changes in customer demands would likely translate into changes in internal priorities.

Large business customers might have higher leverage to make demands. Esp customers that channel a lot of demand, like API wrapper startups (eg Cursor), could make strong requests to AI companies.

Lastly, the government could exert pressure when giving large contracts eg using AI for defence. As a starter, governments could refrain from giving huge defence contracts to AI companies whose new model just [went on a pro-Nazi spree](https://www.theguardian.com/technology/2025/jul/14/us-military-xai-deal-elon-musk). More ambitiously they could attach stringent safety requirements and information security for models used in the military.

*Who?* individual users, business customers, governments

## Shape their market

Companies are responsive to market pressures. Competitors have leverage by shaping the market they both partake in. For example one company might release models with new capabilities, thus putting pressure on other companies to catch up and release similar models. Other market shaping actions include race dynamics to take a market or reach a capability first, pricing strategies, release strategies (eg open-source AI commodifying chatbots), customer acquisition strategies, adding new features or making specific aspects (e.g. safety) more salient to customers.

*Who?* Competitors

## Change their talent pipeline

Another critical input into the AI production function are the people developing and deploying AI. They have a rare and highly-sought skillset without which the companies can’t survive. This was shown in Meta’s recent spending spree to [buy top-talent](https://youtu.be/7OseZlldvYI). This gives candidates leverage to make demands from their future employers. Furthermore, it’s a vector for competitors to influence each other. For example Anthropic poached large parts of OpenAI’s alignment team. This brain-drain hurt OpenAIs public image wrt safety.

Others can attempt to influence the makeup of the talent pool of AI companies. For example field-building programs like MATS aim to produce and accelerate people that care about AI Safety and could end up working at AI Companies. This increases the percentage of candidates that care about safety. Top universities are also in a position of power by educating students about the importance of ethical AI development.

*Who?* potential employees, field builders, educators, competitors

## Insure them

Companies are often required to or benefit from being insured. For example D&O insurance is required by board members and Professional liability can be required to sell to enterprise customers. Insurance companies are experts in risk assessment and could thus play an important role in setting prices for and requiring certain safety practices. For example Lloyd's of London could state that large AI companies' products need safety certification to be insured.

*Who?* Insurance companies

*Thanks for feedback to Claude.*

*I made some memes for this post, but decided they didn’t fit the vibe.* [*Click here*](https://drive.google.com/drive/folders/1HVR5JQ-XSix5McIzOQuodAAWnd-fQZbd?usp=sharing) *if you still want to see them.*

# Impact Analysis of Levers *[This will go into a follow-up post]*

In a previous post I outlined 16 vectors for influencing AI companies, but left out any analysis of their impact. I will now go ahead and make up some numbers.



Specifically I will use an adaptation of the classic Scale, Tractability, Neglectedness framework to estimate the marginal impact of influencing these levers.

1. Scale: How much influence does this lever have over safety relevant decisions in the company?
2. Tractability: How much can altruistically minded people influence this lever? For example insurance companies have profit and regulatory constraints on them and thus can’t be moved as easily.
3. Neglectedness: How many resources in AI Safety are already going into this?
4. Marginal Impact = Scale x Tractability x Neglectedness

This is more of an exercise meant to demonstrate a way to estimate the impact. Also I’m hoping to trigger people (maybe my future self?) into making a more serious attempt at this. **Please don’t take any of these numbers seriously or god forbid even base a decision on them.**

| Intervention | Scale | Tractability | Neglectedness | Marginal Impact |
| --- | --- | --- | --- | --- |
| Join them |  |  |  |  |
| Build something for them |  |  |  |  |
| Collaborate with them |  |  |  |  |
| Regulate them |  |  |  |  |
| Take them over |  |  |  |  |
| Sue them |  |  |  |  |
| Set standards for them |  |  |  |  |
| Convince them |  |  |  |  |
| Spread memes to them |  |  |  |  |
| Setting societal expectations for them |  |  |  |  |
| Invest in them |  |  |  |  |
| Sell to them |  |  |  |  |
| Buy from them |  |  |  |  |
| Shape their market |  |  |  |  |
| Change their talent pipeline |  |  |  |  |
| Insure them |  |  |  |  |