Corporate Orthogonality: Structural Misalignment and the Limit of Technical Approaches

Introduction

Most broadly, I think of AI alignment as the attempt to develop AI systems that ultimately serve human flourishing. What flourishing consists of, I cannot tell you, as its precise contours remain beyond my grasp. I can, however, identify clear instances of its absence and observe corporations systematically producing this absence. The need to avoid such outcomes through careful design dates to the field's origins. In 1960, Norbert Wiener made clear the issue now thought of as the Alignment Problem:

> *If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it… then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.*[1]

More recently, Nick Bostrom, in his 2014 book *Superintelligence*, posited two theses that have guided much of AI alignment research over the past decade. **The Orthogonality Thesis** states that intelligence can be combined with any final goal, such that high capability does not imply benevolence, and the **Instrumental Convergence Thesis** states that intelligent agents with diverse final goals tend to converge on specific instrumental sub-goals like self-preservation and resource acquisition.[2] Through these, Bostrom made clear that without particular effort to align these systems, they may pursue objectives harmful to human flourishing and existentially threatening. This paper concerns the former: the systematic susceptibility to compounding harms orthogonal optimization produces harms that may not end human life, but will reliably diminish its quality.

Consequently, the issue of aligning increasingly capable AI systems has largely been posed as an engineering challenge: how do we technically build systems whose objectives align with those of humans? This framing has culminated in research programs focused on methods such as Reinforcement Learning from Human Feedback, Constitutional AI, and Mechanistic Interpretability. This technical framing, however, often neglects critically normative dimensions of the Alignment Problem: What values? Whose? Implemented through which institutional means?

Before proceeding, it is useful to distinguish between three senses of “alignment” that are often conflated.

---

[1] Wiener, Norbert. "Some Moral and Technical Consequences of Automation." *Science*, vol. 131, no. 3410, 6 May 1960, pp. 1355–58.

[2] *Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford UP, 2014.*

- Technical Alignment: When AI systems reliably do what their developers intend. Following instructions, avoiding unintended behaviors, operating within specified constraints.
- Value Alignment: When AI systems produce outcomes that are good for humans and whether their internal states reflect the same values that we hold, whether it is conducive to human flourishing.
- Institutional Alignment: When AI developers have optimization targets that reliably track human flourishing and when such structures select for beneficial outcomes.

Technical alignment is necessary, but not sufficient, for value alignment. Further, value alignment remains out of reach so long as institutional pressures systematically preclude the research, resources, and deployment decisions it requires. The central concern is how the principal institutions in AI development undermine both technical and value alignment efforts.

A fatal assumption persists: that the structures through which AI is developed are neutral vessels for product development and research, that we have a sufficient level of institutional alignment through which value alignment is tractable.

This assumption is false. Today, the principal structural element of the AI ecosystem is the corporation. It is the primary organization through which everything from applied product development to frontier research takes place. However, the corporation is not a benign static layer on top of which research is conducted; it is itself a goal-directed system optimizing for a specified objective function. I seek not to claim that corporations are intelligent systems in any proper sense, rather that they, as systems, exhibit the exact structural features that alignment researchers identify as dangerous in AI systems. Namely, the corporation is subject to features of narrow goal pursuit, orthogonality, and instrumental convergence.

Viewing the corporation through this lens reveals a recursive element to the alignment problem. We are attempting to build aligned AI systems within institutions that themselves exhibit the structural features of misaligned agents. The technical success of any alignment research remains downstream of, and wholly dependent on, a misaligned substrate. This paper argues that the structural dynamics create conditions for recursive alignment failure. Attempts to align AI within systematically misaligned institutions face persistent pressure toward ultimate misalignment. These dynamics culminate primarily in a systemic underinvestment in alignment research and in an incentivization to deploy products hostile to human flourishing. This is uniquely urgent in the AI industry due to the speed of development, integration into critical infrastructure, and general opacity that compound its risks far beyond other forms of corporate misalignment.

Technical alignment researchers worry that using an imperfectly aligned AI to oversee the training of a more capable AI may compound alignment errors across successive generations.

We, however, fail to apply this same logic to the base model of the entire chain: the corporation. If the base model providing the initial oversight, resource allocation, and deployment decisions is itself misaligned, the recursive chain is poisoned from the start.

In the following, I will first establish the grounds for analyzing corporations as systems exhibiting misaligned optimization dynamics, drawing on Bostrom's theses as diagnostic tools. I will then turn to examine the mechanisms by which this institutional misalignment manifests, which I call the problems of omission and commission, arguing that achieving value alignment is ultimately constrained by a lack of institutional alignment.

Foundational Substrates of Misaligned Optimization

A Functional Analogy to Intelligent Systems

Corporations exhibit many of the functional features that generate alignment concerns in AI systems. While I do not claim corporations to be "intelligent" in any philosophically robust sense, they display importantly analogous behaviors of:

> Goal Direction: Corporations operate in service of a specified objective function; namely, they bear a fiduciary duty to maximize shareholder value.
> Information Processing: Corporations take in, process, store, retrieve, and use information to understand parameters of their environment.
> Learning/Adaptation: Corporations update their internal states based on environmental information and adjust subsequent decisions based on these findings.
> Planning, Reasoning, Problem-Solving: Corporations coordinate across their constitutive parts through complex processes to arrive at solutions that further their objective function.

These features make corporations powerful optimizers. They command resources sometimes exceeding those of nations, operate over time horizons far exceeding those of individual human agents, and coordinate highly complex systems toward unified objectives. In terms of optimization capacity and systemic complexity, corporations resemble what AI researchers envision as superintelligent systems, operating at levels no individual human can fully comprehend or control.

Corporate Orthogonality and Instrumental Convergence

Bostrom's Orthogonality Thesis states that "intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal."[3] An intelligent system can be arbitrarily capable while pursuing any goal. High capability is not

[3] Bostrom, p. 130.

coupled with values that meaningfully overlap with those of humans. This idea grounds much of modern alignment research: we cannot assume increasingly capable AI will naturally share human values.

Corporations are subject to the same structural concerns. A corporation can be maximally capable of achieving its objective, while that objective remains orthogonal to human flourishing. This manifests as specification gaming.[4] Maximizing shareholder value originated as a proxy for social welfare, but once this measure became the target, in congruence with Goodhart's Law, it ceased to be a good measure.[5]

We observe corporations successfully optimizing for the proxy of profit, even when it actively decouples from the true target of flourishing, and there are clear examples of this in the past: the opioid epidemic, the financial crisis, and, more recently, the engineering of polarized and addictive social media. These are not bad actors with malicious intent but predictable outcomes of powerful systems subject to goal misspecification.[6] These corporations are remarkably successful optimizers but of a misspecified objective function.

Furthermore, corporations display convergent instrumental sub-goals that Bostrom predicts for advanced AI: self-preservation through lobbying and regulatory capture; goal-content integrity in rigid adherence to shareholder primacy; and resource acquisition through attention capture and natural resource exhaustion, among others. Instrumental convergence is not inherently pernicious; however, when these instrumental behaviors serve an orthogonal final goal, their pursuit often proves harmful in effect. We thus see paved paths by which the dominant institutions of AI development might inhibit adequate alignment and incentivize the deployment of technologies misaligned with human flourishing.

The Limits of Legibility

The corporation-AI analogy breaks in one critical respect: legibility. Neural networks remain opaque "black boxes," whereas corporations are structurally transparent and decomposable. We can, in principle, examine their decision-making processes, trace their incentive structures, and identify their optimization targets. This transparency ostensibly offers far greater opportunity for control than we have with AI.

---

[4] "A behaviour that satisfies the literal specification of an objective without achieving the intended outcome." See www.deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/ for more.

[5] Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." *European Review*, vol. 5, no. 3, 1997, pp. 305–21.

[6] "Occurs when the specified reward function does not accurately capture the true objective or desired behavior." This is importantly distinct from specification gaming in that it is the designer of the reward function, not the AI system, that "misspecifies" what is to be optimized for. See www.lesswrong.com/posts/mMBoPnFrFqQJKzDsZ/ai-safety-101-reward-misspecification.

However, despite this legibility, we consistently fail to prevent corporate harms. The Sacklers and Enron executives were held accountable only after severe, lasting damage was done, revealing something of critical importance: alignment failures are not just epistemic, they are political. We can see how and why corporations pursue harmful paths, yet we have consistently failed to construct adequate constraints. Even with full interpretability, we would face the same political and institutional obstacles that have prevented corporate alignment. Understanding misalignment is not the same as having the institutional capacity, or will, to correct it.

This should further temper our optimism about AI alignment. If we struggle to align systems whose internal processes and constitutive parts are intelligible, how do we expect to align a system whose processes and parts are not? The challenge of alignment only compounds with systems we cannot interpret. Legibility may be helpful, but it is not sufficient for alignment, and we don't even have that advantage with AI. With this in mind, we can now examine the specific mechanisms by which institutional misalignment manifests.

Mechanisms of Recursive Failure

Faulty Foundations

The outcomes of alignment research depend critically on the structures in which they are created. A pervasive fallacy holds that such structures are neutral mechanisms granting researchers autonomy; in reality, work passes through selective filters of resource allocation and financial viability. An analogous process occurs in academia, and this is not an argument for relocating AI research, but for recognizing the current pressures that shape it.

The structural realities of the AI industry lead to a systematic underinvestment in alignment research—a fundamental problem of omission. Clearly, alignment research is done. But this is despite the structural impediments, not because of them. The argument here relies upon five premises from which a natural conclusion is drawn:

> Premise 1: AI companies operate with finite resources.
> Premise 2: Corporate incentive structures reward allocation of resources toward activities with clear financial returns.
> Premise 3: Technical alignment research is a cost center, not a profit center.
> Premise 4: Given corporate orthogonality, there is no robust mechanism coupling profit-maximization to investment in alignment research.
> Premise 5: AI alignment depends on adequate investment in technical alignment capabilities.
> Conclusion: Corporate incentive structures systematically produce underinvestment in alignment research.

Firstly, AI companies, like all other corporations, operate with limited resources. Human labor, compute, and energy are well-known constraints on AI development, and companies reliably allocate these scarce resources toward activities with the highest expected financial return. Further, technical alignment research is perceived as a cost center. Alignment efforts do not translate directly into revenue, they can slow development or discourage deployment, and their direct contributions to final commercial goals are often diffuse and opaque.

On the other hand, it is sometimes tempting to believe technical alignment is sufficiently linked to a system's utility. To a limited extent, this is true: a degree of basic alignment is intertwined with capability, as an AI that fails to perform the task or follow instructions will have significantly lower utility. But this line of thinking misses two critical distinctions.

The first is between short- and long-term risk. Deep alignment, ensuring a system remains safe under capability gains, produces returns that are distant and diffusely distributed across society rather than captured by the firm. Discount rates systematically underweight such investments relative to those related to near-term capability gains.

The second is between a thin and thick conception of alignment. Thin alignment ensures a system behaves as intended, that it follows instructions and avoids obvious harms. Thick alignment ensures the system is genuinely aligned, that its objectives and internal states truly reflect human flourishing. Each step in AI development is like adding a floor to a skyscraper. There are visible ways each floor must be sound—framing, joists, support beams. But the higher we build, the more critical the hidden elements become: foundation depth, rebar strength, footing integrity. The minimal alignment intertwined with capability is like ensuring each floor stands, and it distracts from the crucial task of reinforcing the foundation itself. If we do not lay a foundation designed for the weight of superintelligence, the skyscraper risks collapse, taking both itself and the surrounding community down with it.

This dynamic reveals the core problem of corporate orthogonality: there is no robust mechanism directly coupling profit-maximization to investment in deep, long-term alignment. One might object that reputational concerns, regulatory anticipation, or liability exposure provide such a mechanism, and while these couplings do exist, they are weak. Reputational concerns require the appearance of alignment, not alignment itself. Regulatory anticipation addresses only a fraction of alignment concerns, those which regulatory bodies can foresee, identify, and measure. Liability often only applies after harm has materialized and to a minimal degree. These mechanisms select for levels of superficial alignment sufficient to satisfy external observation and not for the deep structural alignment that long-term safety requires. The gap between what is incentivized and what is needed remains vast and unaccounted for.

Along this path, an AI developer can be maximally successful by every market metric while investing minimally in deep AI safety, and they are in fact structurally incentivized to do so. If each additional floor built seems structurally sound with its superficial alignment, and reassessing the foundation incurs additional costs and time delays, the incentive is clear.

Add on top of this the competitive dynamics of the AI race—even if a company might otherwise accept the additional costs and time delays of reinforcing their foundation, a competing firm might in the meantime finish building their next floor. With this, our company then bears a practical obligation to rush to completion, prioritizing the immediate, revenue-generating act of opening the next floor to tenants while ignoring catastrophic foundational weaknesses. But this is not just an incentive problem, it is a collective action problem. Even if every firm preferred a methodical assurance of long-term, thick alignment, no firm can unilaterally do so without risking elimination. Race dynamics preclude scenarios all parties might otherwise prefer.

Alignment in service of flourishing necessitates profound investments, not just because of its importance, but its difficulty. Rather than pausing to lay this essential, deep bedrock that the true weight of a superintelligence demands and iteratively reinforcing it, corporate incentive structures command resources for short-term value creation—to build the next floor.

Again, this is not an idiosyncratic or contingent failure of a particular company or set of companies or an accusation of malicious intent. Instead, this is a predictable consequence of powerful optimization systems operating under misaligned targets. The most significant institutions at work in the development of AI have structurally disempowered us from dedicating the necessary time and resources to alignment research.

We have seen this very phenomenon unfold with OpenAI: the company was originally started as a nonprofit with a comprehensive corporate charter. It then turned to a capped-profit model where the nonprofit arm still held significant governance influence, then most recently to a Public Benefit Corporation whose board has sought to remove the nonprofit's influence entirely. With this and the increasing capitalist incentive structures at play in the organization, we've seen a reduction in alignment-focused research, the dissolution of their Superalignment team, and an exodus of key alignment researchers. But the story of OpenAI is not one of corruption or greed; it is a story of structural pressures that marginalize alignment efforts and safe AI deployment. Through a gradual shift in the company's optimization target, we saw a parallel shift away from alignment research and are likely to see the deployment of increasingly harmful products. The corporation operates as a highly efficient optimizer whose pursuit of its target dominates anything that might stand in its way, including the company's stated mission or the goals of any individual constitutive agents.

The Problem of Commission

While the problem of omission concerns what corporations fail to do, the problem of commission regards what they are compelled to do. As orthogonal corporate incentive structures continue to dominate AI development, we will predictably see an increase in the deployment of AI systems that are actively hostile to human flourishing. This stems from a core alignment failure: maximizing shareholder value is an intensely lossy conversion of genuine human wellbeing. Further yet, corporations can be seen to converge on instrumental behaviors like market expansion and attention capture, while remaining limitedly accountable for negative externalities. Crucially, companies do not merely lack an incentive to build properly aligned AI systems; they are structurally incentivized to pursue harmful, yet lucrative, deployments. We might formalize this pitfall of institutional structure in the following way:

> Premise 1: Corporations converge on instrumental behaviors to maximize their objective function.
> Premise 2: For-profit companies face limited accountability for diffuse negative externalities.
> Premise 3: There exist profitable deployments of AI that are hostile to flourishing.
> Conclusion: AI companies will predictably pursue harmful deployments when they represent an efficient path to the objective function.

The third premise may seem obvious, but it warrants examination. The question is not whether harmful AI deployments exist, as history provides an abundance of examples. The question is whether their profitability is contingent or structural. I contend it is structural: in an underconstrained environment, harm and profit overlap systematically wherever value extraction is more efficient than value creation, wherever negative externalities can be offloaded onto society rather than absorbed by the firm.

We are currently living through this dynamic via the attention economy. What began as a seemingly benign effort to accrue daily active users soon evolved into sophisticated addiction engineering.[7] Despite documented harms (psychological dependence, polarization, and misinformation), platforms remained misaligned because the objective function was pursued via an instrumental goal of engagement. Because the negative implications remained externalized, borne by society, the system continued to optimize, creating more harm. We've seen similar dynamics play out across industries, and we have no reason to believe this dynamic will stop with AI—In fact, more capable systems might just be that more capable of these very extractions. We are already seeing groundwork for such extractions with AI, exemplified by

[7] See Eyal, Nir. *Hooked: How to Build Habit-Forming Products*. Portfolio/Penguin, 2014.

Mark Zuckerberg's vision to own 80% of our friends, monetize intimacy, and manipulate emotional vulnerability as a vector for extraction.[8]

Critically, though, this does not require malice. I am certain industry leaders intend to build products that improve the world, and I further remain convinced that a lot of what they build will achieve that. However, the intentions of these leaders remain downstream of a greater structure that demands the pursuit of an orthogonal goal, and the positively valenced products they do deploy must ultimately serve this goal. Harmful products often are not the result of malicious actors, but of an intense optimization pressure the system exerts on its constitutive parts. If an AI product can be monetized through exploitation, and if the corporation must maximize profit, without the necessary side constraints in place, the system will do so. Most simply, it comes down to a design flaw of misspecification: the system is working exactly as designed; the design is simply misaligned.

Conclusion

In our efforts to develop AI systems which act in service of human flourishing, technical alignment may be necessary, but it is certainly not sufficient. The structures in which AI is being developed have proven themselves orthogonal to our aims with biases towards instrumental means which often counter our very goals. Solving the technical problem of alignment while ignoring the broader misaligned ecosystem will yield sophisticated tools for ultimately misaligned ends. We are unlikely to achieve value alignment via Constitutional AI while our efforts remain trapped in institutionally misaligned systems. The structural misalignment at hand is itself a first-order alignment problem that needs to be immediately addressed.

This is in no way an indictment of AI companies or people within them. Intentions are largely pure, but the relative irrelevance of intentions is precisely the point. When enveloped by intense systemic pressures, constitutive efforts matter little in changing the optimization process or goal. The story of corporate AI development is not one of villainy, but one of systems pursuing misspecified objectives.

It is neither one of hopelessness. Neither corporations nor their objective functions are natural inevitabilities—they are legal and social constructs, constructs which we collectively hold the power to change. The limited liability corporation is a recent invention, shareholder primacy more recent still. While the development of harmful products may be structural, the structures by which it is so are themselves contingent. They can be changed, but only if we recognize them as the first-order problem they are. The corporate form itself is not the problem. Corporations long predate the notion of shareholder primacy and were originally chartered to serve specific public

[8] "Mark Zuckerberg – AI Will Write Most Meta Code in 18 Months." *Dwarkesh Podcast*, hosted by Dwarkesh Patel, 29 Apr. 2025, www.dwarkesh.com/p/mark-zuckerberg-2. In the interview, Zuckerberg discusses Meta's AI companion strategy, suggesting that AI agents could eventually constitute the majority of users' social relationships.

purposes. Shareholder primacy is a relatively recent invention, a product of 20th century neoliberal economic theory. It is not a natural law; it is a design choice. Just as the problem with AI isn't the technology itself but the specifications and fulfillment of its objectives, the problem with corporations is not the corporate form but its misspecified objective function.

One might object that alternative corporate structures already exist, ones that do not serve the notion of shareholder primacy (public benefit corporations, nonprofits). The issue is that while such constructions do provide partial mitigation, their differences erode over time due to structural pressures. OpenAI's trajectory from nonprofit to capped-profit to public benefit corporation illustrates precisely this erosion. Alternative structures may help but are not sufficient without broader changes to the competitive environment. Regulation through behavioral constraint alone is insufficient for adequate value alignment. Structural reform to the objective function of the corporation, or a relocation of frontier AI work outside of profit-maximizing structures, is likely necessary for the development of AI that reliably serves human flourishing. That being said, institutional alignment does not answer the deeper normative question of whose values AI should serve. It merely makes answering that question possible.

It is worth emphasizing further why this matters more for AI. Intelligence is not a discrete product confined to its own industry—it is infrastructure that will integrate itself into every institution and the very rhythms of daily life. Further, the speed and scale of its development far outpaces both regulatory and social adaptation, and its integration is not easily reversible. Once it is woven in, dependencies become structural. The depth and irreversibility of AI integration means that misalignment at the institutional level will propagate throughout society and in ways that are difficult to anticipate and harder to undo. Taken with AI’s opacity and recursive capability gains, the urgency of addressing institutional misalignment becomes clear.

The alignment community is right to worry about orthogonality, but it has yet to reckon with the full implications of it. If a capable system may pursue any objective in principle, and if the superstructure developing such a system is itself a capable optimizer subject to orthogonality, then the efforts to align AI are inheriting a misaligned prior. Without the proper foundation, technical efforts remain all but futile, not because they are poorly executed, but because they are embedded in institutions that neither reward nor reliably implement their findings. Institutional alignment is not an alternative to technical alignment but a precondition for it to really matter. Technical alignment work will remain sidelined if the institutions deploying AI do not have the right objectives. We cannot align AI downstream of a misaligned substrate and expect aligned outcomes. Orthogonality is not only a problem for artificial intelligence, and the question is whether we will recognize this in time to do something about it.

Works Cited

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford UP, 2014.

Eyal, Nir. *Hooked: How to Build Habit-Forming Products*. Portfolio/Penguin, 2014.

Krakovna, Victoria, et al. "Specification Gaming: The Flip Side of AI Ingenuity." *DeepMind Blog*, 21 Apr. 2020, www.deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

"Mark Zuckerberg – AI Will Write Most Meta Code in 18 Months." *Dwarkesh Podcast*, hosted by Dwarkesh Patel, 29 Apr. 2025, www.dwarkesh.com/p/mark-zuckerberg-2.

markov. "AI Safety 101: Reward Misspecification." *LessWrong*, 18 Oct. 2023, www.lesswrong.com/posts/mMBoPnFrFqQJKzDsZ/ai-safety-101-reward-misspecification.

Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." *European Review*, vol. 5, no. 3, 1997, pp. 305–21.

Wiener, Norbert. "Some Moral and Technical Consequences of Automation." *Science*, vol. 131, no. 3410, 6 May 1960, pp. 1355–58.