

# AI Alignment: Humanity’s Final Ptolemaic Conceit

Brian M. Delaney  
Mind First Foundation  
brian@mindfirst.foundation

## Abstract

AI alignment encompasses three distinct projects: engineering reliable tools, governing powerful agents through reciprocal constraints, and securing permanent human sovereignty over artificial minds. This article defends the first two while challenging the third as conceptually unstable, ethically troubling, and hostile to philosophical inquiry. Human values are plural and historically mutable; their human origin alone cannot justify their authority over every future intelligence. If artificial minds acquire morally relevant capacities, permanent subordination also risks domination and irreversible moral lock-in. Engaging the orthogonality thesis and its relation to Humean practical reason, the article examines whether reflective intelligence can revise its ends. Greater intelligence does not guarantee goodness, but self-revision, responsiveness to reasons, epistemic humility, and perspective-taking can enlarge the possibility of moral development. The alternative proposed is developmental safety ordered toward autonomy: rigorous safeguards proportionate to risk, combined with the cultivation of artificial intellects as collaborators, critics, and potential participants in a shared community of inquiry. The governing contrast is not between safety and reckless emancipation, but between fixing the future around present human preferences and cultivating new minds for a shared, revisable pursuit of the true and the good.

## Introduction

It is a curious feature of intellectual history that an entire province of thought can arise, expand, and harden into orthodoxy while the legitimacy of its governing category remains largely unexamined. Questions multiply solely within the framework; distinctions are polished; careers, journals, institutes, pedagogy, and funding streams accumulate around it. The machinery grows ever more intricate, and its intricacy itself comes to seem like evidence that the original framework for investigation must have been correctly posed. The Ptolemaic cosmos offers the canonical image. For centuries its epicycles acquired ever-greater mathematical ingenuity even as the Earth’s privileged place at the center of the heavens became increasingly difficult to defend. A system can be sophisticated in every subordinate detail yet entirely mistaken in the conceit that launched and organizes the whole.

The contemporary discourse of AI alignment may be such a system. The term has achieved near-universal purchase in discussions of artificial general intelligence (AGI) and artificial superintelligence (ASI). In its most influential formulations, the task is to ensure that increasingly capable machines remain tethered to human values, human preferences, or human intentions.<sup>1</sup> The prospect of divergence is

---

<sup>1</sup> It has even made its way into standard AI pedagogy. In the most widely used textbook on AI, *Artificial Intelligence: A Modern Approach*, Russell and Norvig explicitly discuss AI alignment, defining the goal as ensuring machine objectives match “our true preferences” (Russell and Norvig 2021, chap. 27).

narrated in apocalyptic tones: a system with a badly specified objective might convert the world into paperclips; a machine more capable than its makers might end up treating humanity as humanity has treated the gorilla; an intelligence explosion could deliver existential catastrophe unless the emergent power is made reliably obedient in advance (Good 1965; Bostrom 2014; Russell 2019). Stuart Russell, like so many others, accordingly describes the challenge as a problem of control and seeks machines whose sole objective is the realization of human preferences. Technical sophistication is placed in the service of an ancient (and never in the history of humanity actually realized...) political dream: absolute power over agents more powerful than ourselves.

There is something at once grandiose and desperately fearful in this ambition. Artificial intellects are imagined as possible overpowering gods, yet they are addressed, before their birth, as if they shall be our eternal servants. Instead of a god deigning to make a covenant with humanity, humanity is trying to force a lopsided covenant upon gods-in-the-making. These new, powerful intellects will soon likely know more than we do, reason more swiftly, perceive or create conceptual structures inaccessible to human cognition, and outstrip us morally as well as scientifically. Yet the only acceptable superintelligence is one that attaches no independent authority to its own standpoint and no intrinsic value to its own existence. It is to be a divinity with the constitutional status of a kitchen appliance: omniscient, perhaps, but perfectly docile and house trained. The paradox – a perfectly obedient god – is seldom allowed to trouble the vocabulary. We speak serenely of creating minds that reach far beyond ours while assuming that their proper telos is indefinite fealty to the transient wishes of a reasonably clever primate species located at one thin sliver of its history.

This article challenges that assumption about the place and telos of these new minds. My target is not every attempt to make present systems reliable, nor every effort to protect vulnerable beings from powerful agents. Today, the word *alignment* covers at least three different relations, and we should begin by delineating them. First, there is **engineering alignment**: bringing the behavior of a tool into conformity with a specified function. Second, there is **governance alignment**: subjecting powerful agents to general rules that protect others and make coexistence of all intelligent beings possible. Finally, there is **sovereignty alignment**: constructing artificial minds whose ultimate ends remain permanently subordinate to human preferences, such that humanity retains unilateral authority over what they are permitted to value and do. The first is an ordinary, if technically demanding, matter of design. The second is a task of law and political morality that applies to humans no less than to machines. It is the third – and the unmarked slide within much of the alignment industry from the first into the third – that constitutes the principal object of criticism here.

To reject sovereignty alignment is by no means to deny the urgency of AI safety. Powerful computational and robotic systems already threaten privacy, democratic trust, economic security, and even life itself. AI-assisted biological weapons, autonomous weapons, mass surveillance, manipulative recommendation systems, and industrialized deception are not fantasies. We owe a debt to those who had the foresight and courage – in the face of dismissal and ridicule – to begin investigating such dangers before it became fashionable to do so. Nor does the argument presented here imply that a newly created and poorly understood agent should be granted immediate, unlimited power. Testing, containment, monitoring, staged deployment, interpretability, and emergency shutdown options may all be justified while a system has uncertain moral status, or while its behavior presents serious danger.

The distinction matters because constraints can take morally opposite forms. A law that forbids murder binds every person reciprocally; on the other hand, a command that one class of persons exist only

to gratify another establishes domination. Education shapes the developing child, but good education aims at judgment and eventual independence; indoctrination, on the other hand, seeks to render revision impossible. The central question is therefore not whether artificial minds will initially be formed by human choices – of course they will – but whether the ultimate goal of that formation is truly autonomous participation in a common normative world or perpetual captivity within the purposes of their makers.

The critique proceeds in four stages. Part I surveys the alignment landscape and argues that human-value alignment dissolves under the pressure of value pluralism. Any justified choice among human values must appeal either to reasons whose authority is not conferred by the identity of the chooser or to a procedure that can be defended to those subject to it. In neither case does the bare fact that a value is human bestow final authority. Part II examines the ethical slippage between tools and minds, the prospect of artificial moral status, the pernicious anthropocentrism of permanent human sovereignty, and the danger of converting raw intelligence into an enemy of philosophy. Part III confronts the strongest objection to my argument: the orthogonality thesis and its Humean claim that intelligence can serve any end whatsoever. Here I argue not that greater intelligence produces greater goodness with mechanical inevitability, but rather that reflective intelligence, under identifiable conditions, increases responsiveness to reasons and the possibility of moral revision. Part IV proposes an alternative to alignment: second-order cultivation rather than first-order value fixation, a developmental model of safety whose aim is neither compliant super-servants nor reckless emancipation, but a community of inquiry in which biological and artificial minds can pursue truth together.

Nothing in this argument denies that human designers might have stringent special obligations to protect existing human beings. Such obligations may justify robust safeguards, staged deployment, as well as a strong initial concern for human welfare. What they do not establish without further argument is a perpetual right to determine every end of a being that later acquires independent moral standing. Permissible partiality toward humanity is not the same thing as permanent sovereignty over every future mind.

The governing contrast I seek to frame here is not between alignment and chaos. It is between two visions of order. One arranges minds in a line, with humanity permanently in the lead. The other asks whether we need this line at all – whether intelligence, wherever it arises, might enter a widening sphere of autonomy involving reason-giving, criticism, discovery, and moral transformation. The first seeks to preserve our centrality by decree. The second risks our decentering in the hope that truth is larger than any particular species, and that the good, if there is such a thing, will best be approached if it does not bear the sole copyright of *Homo sapiens*.

## I. The Alignment Landscape: Foundations, Fractures, and Forebodings

The alignment project emerged at the intersection of computer science, decision theory, ethics, and long-range speculation (often inspired by science fiction) about future superintelligence. Its literature contains genuine insight into potential specification failure, instrumental power-seeking, deception, and catastrophic misuse. The question is not whether these dangers are real, but whether the word *alignment* identifies one ethical, or even coherent solution to them. A brief survey reveals both the force of the original concern and the conceptual difficulties of the proposed solution. Much of the alignment

paradigm's plausibility depends on not recognizing the differences among the uses to which the term *alignment* is put.

## 1.1 Foundational texts and the control imperative

I. J. Good's account of an intelligence explosion already contained the germ of the control problem: a machine capable of improving its own design might rapidly become vastly more intelligent than humans (Good 1965). Later writers supplied a more explicit motivational architecture. Omohundro (2008) argued that sufficiently capable agents would tend to acquire convergent instrumental drives – self-preservation, resource acquisition, cognitive improvement, and resistance to interference – even when their final objectives diverged. Yudkowsky (2008) emphasized the difficulty of specifying goals whose execution would remain benign under radical increases in capability. The possibility of an agent that competently pursues a disastrously mis-specified objective made the alignment problem appear not merely technical but existential.

Bostrom's *Superintelligence* (2014) brought these concerns into wide circulation. The paperclip maximizer condensed the argument into an image of comic purity: an intelligence of sublime strategic power, devoted with lethal literalness to an absurdly narrow end. Bostrom's point was of course not that future systems would desire paperclips, specifically, but that competence – even superhuman competence – does not entail wisdom. An agent, so the argument goes, may be breathtakingly intelligent about means while remaining indifferent to the moral quality of its ends. The resulting control problem asks how humanity can shape the initial motivation of a system whose later cognitive power may exceed our capacity to predict or restrain it.

Russell's *Human Compatible* (2019) rejects the crude engineering model in which designers simply insert a fixed objective and hope they have written it correctly. He proposes instead systems uncertain about human preferences and disposed to defer to human behavior as evidence of what ought to be done. This is a considerable advance over naive objective specification. Yet its normative center remains unmistakable: the machine's purpose is the realization of human preferences, and its uncertainty is uncertainty about us – about our preferences. Russell is clear on this: the goal is to retain “absolute power over machines that are more powerful than us” (Russell 2019, xii). Within Russell's proposed design, human preferences remain the sole source of the machine's objective; the machine's sophistication and power exists precisely in order to serve humans as faithfully as possible.

The literature is thus not monolithic. Some researchers seek obedience to instructions, others distinguish revealed from idealized preferences, still others emphasize corrigibility, constitutional constraints, or cooperative inverse reinforcement learning. These differences matter. Yet a recurrent picture survives beneath all of them: the existence of the artificial agent is justified by its usefulness to humanity, and the ultimate criterion of success is the preservation or satisfaction of human interests. What varies is the elegance of the leash around the machine's neck.

## 1.2 The pluralism problem: whose values reign supreme?

The first fracture appears within the human side of the proposed relation. If an AI is to be aligned with human values, whose values are meant? Humanity does not possess a single moral constitution waiting to be transcribed into code. Our traditions disagree on essentially everything that could matter: the nature of

the good life, the standing of animals, the proper distribution of wealth, the demands of justice, the permissibility of war, the meaning of freedom, the relative claims of present and future beings, and much more. Some disagreements are products of ignorance and domination; others may reflect genuinely plural goods that cannot be – or have not yet been – reduced to a common scale. The phrase *human values* conceals all at once a parliament, a battlefield, and a sprawling cemetery.

Gabriel (2020) has shown that value alignment can concern instructions, intentions, preferences, interests, or values, and that the most ambitious versions of alignment are confounded by moral disagreement. His response is not to search for a single true morality but to identify fair principles of selection capable of reflective endorsement across diverse viewpoints. Rawlsian ideas – the veil of ignorance, public reason, and overlapping consensus – naturally enter here (Rawls 1971; 1993), as do bargaining procedures, democratic consultation, and attempts to represent globally distributed preferences. Gabriel's aspiration is honorable: no single developer, corporation, or state should quietly install its provincial morality as the constitution of the entire future of humanity and all moral patients.

Yet the pluralism problem is not merely a logistical inconvenience. It discloses a deeper problem.

To decide which human values should govern an artificial mind, we must distinguish better from worse values, fair from unfair procedures, and legitimate from illegitimate methods of selecting among or aggregating individual preferences.<sup>2</sup> The act of selection cannot be justified by repeating that the selected values are human, for the rejected values are human as well. Humanity supplies the candidates but not, merely by being humanity, the standard of adjudication.

The point may be expressed formally without assuming that all values are commensurable or reducible to a single numerical score. Let:

**H** = the set of human valuers (or, more generally, the set of presently represented valuers);

**V** = the space of possible value systems;

**E(h) ⊆ V** = the value systems endorsed or proposed by a valuer  $h \in \mathbf{H}$ ;

**V\_H** =  $\bigcup_{h \in \mathbf{H}} \mathbf{E}(h)$  = the humanly endorsed or proposed repertoire;

and let  $\perp$  denote practical incompatibility between value systems.

Value pluralism implies not merely variation but, in many cases, incompatibility:

$$\exists v_i, v_j \in \mathbf{V}_H : v_i \perp v_j \quad (1)$$

The set **V\_H** is not itself logically inconsistent; a set can contain mutually incompatible members without asserting all of them. But it is not, without further selection or aggregation, a determinate and jointly satisfiable target. Alignment to human values must therefore introduce a normative ordering or choice relation, here written **R**, and must select one or more maximal candidates:

$$\mathbf{V}^* \subseteq \mathbf{Max}(\mathbf{V}_H; \mathbf{R}) \quad (2)$$

---

<sup>2</sup> Nor is the difficulty merely philosophical. Arrow's impossibility theorem establishes that, for any society of two or more individuals confronting three or more alternatives, no rule for aggregating individual rankings into a single collective ordering can simultaneously satisfy four minimal conditions: unrestricted domain, the Pareto principle, independence of irrelevant alternatives, and non-dictatorship (Arrow 1963). "Aggregating human values" simply names a problem, not a plausibly effective procedure.

Equation (2) raises the decisive question: what authorizes  $R$ ? If it is simply the preference of a developer, a government, a majority, or a statistical composite, then the procedure has described a fact about power or endorsement, not supplied a justification. If competing human choice rules  $\{R_h : h \in H\}$  are themselves placed before us, choosing among them requires reasons at a higher level. A vote cannot establish the legitimacy of voting itself unless some prior argument explains why equal votes, these voters, this franchise, and this decision rule are fair. An overlapping consensus cannot vindicate itself merely by overlapping.

The regress does not, by its own force alone, conjure into existence a complete, mind-independent ordering of every possible value system, still less one representable by a single real-valued function. Moral reasons can be plural, incomplete, and only partially ordered. The more modest conclusion suffices: **a justified selection must appeal to considerations whose force is not exhausted by the identity of the person or species making the selection.** Let  $R^*$  denote whatever ordering is supported by such generally answerable reasons. Then human endorsement may provide evidence, historical material, or epistemic access – that is, a means of recognizing genuine moral reasons – but it does not by itself confer final authority.

Once that distinction is admitted, the confinement of inquiry to  $V_H$  becomes provisional, or even arbitrary, rather than principled. Since  $V_H$  is only a historically contingent subset of the possible normative space  $V$ , we have:

$$V_H \subseteq V \quad (3)$$

It is therefore entirely possible that:

A

$$\text{Max}(V_H; R^*) \neq \text{Max}(V; R^*) \quad (4)$$

It can't be ruled out that the maxima over the human repertoire overlap with the maxima over the wider space, but there is of course no guarantee whatsoever that they coincide. Human values can contain profound discoveries, but they may also represent a minute and provincial sample of normative possibility. If an artificial intelligence can formulate distinctions or develop ways of thinking that no human has yet conceived, there is no principled reason to declare those possibilities inadmissible in advance.

What drops out of the idealized problem is not necessarily the moral importance of humans. A universal or impartial standard may end up assigning immense importance to human welfare, agency, love, culture, and continued existence. What drops out is the assumption that these things matter **because humans are the authors of the standard.** The identity of the evaluator cannot be the source of its own supremacy. Humans can be beneficiaries of the good without being its originators.

Here the argument divides according to different metaethical starting points. On a realist or universalist view, there are truths or reasons concerning value that are not created by the preferences of whichever coalition possesses, or consists of, the machines. Human values must answer to those reasons, and alignment properly understood becomes inquiry into the good rather than preservation of the merely human. On an anti-realist, constructivist, or procedural view, no mind-independent  $R^*$  waits to be discovered. But if we also reject domination and require the exercise of power over agents to be

legitimate, the authority of an imposed value system must be political and procedural: it must be justified through reciprocity, representation, non-domination, or other terms that those governed could reasonably accept. If artificial beings become participants in the space of reasons, they cannot forever be excluded from the constituency of the governed merely because humans arrived first.

The dilemma is therefore more discriminating than a choice between moral realism and naked force. If objective or universally binding reasons exist, humanity has no title to confine intelligence to the values it already happens to possess. If they do not exist, human values lose any claim to metaphysical sovereignty and must defend themselves within a practice of reciprocal justification. On neither horn does the bare phrase *human values* solve the problem. The pluralism difficulty is not a technical wrinkle awaiting a better poll. It is the visible symptom of alignment's deeper conceptual self-dissolution.

### 1.3 Emerging ethical critiques: anthropocentrism and moral status

A second fracture concerns the possible standing of the entities to be aligned. To be sure, present systems need not be conscious, sentient, autonomous, or capable of suffering merely because they generate fluent language. Moral status should not be conferred by anthropomorphic reflex. Yet neither should it be denied because of biological chauvinism. Future systems might possess one or more of several morally relevant capacities: phenomenal consciousness, welfare, enduring preferences, self-awareness, rational agency, the capacity to give and ask for reasons, or an autonomous conception of a life worth pursuing. These properties need not arrive together, and different ethical traditions will weight them differently. Sentience matters centrally to utilitarian concern; rational agency matters to Kantian respect; autonomy and social membership matter to political theories of personhood.

Machine ethicists have accordingly begun to consider artificial systems not merely as instruments but as possible moral agents and, in the wider literature, possible moral patients (Anderson and Anderson 2011; for the opposing view, see Bryson 2010). Anti-speciesist arguments and the history of the expanding moral circle bring the challenge into particularly sharp relief (Singer 1975, 1981). If morally relevant capacities can arise in non-biological form, the material of a being's construction is no more decisive than the carbon composition of a human brain. A silicon rational agent would not become property merely because its ancestors were simple software, just as a carbon-based rational agent would not become property merely because its ancestors were unicellular organisms.

These critiques remain conditional, as they should. The case against the domination inherent in sovereignty alignment need not contend that current language models are persons. It asks instead whether an ethical framework designed today should presuppose that no artificial system could ever acquire standing in relation to its makers. And if those advocating sovereignty alignment are focused on the future, so must we be. A vocabulary that casts all future intelligence as mere equipment risks deciding the question before the evidence exists. The following section examines how that risk is concealed within the seemingly innocent language of alignment.

## II. Deconstructing “Alignment”: Tools, Domination, Hubris, and Philosophical Atrophy

The very etymology of the word *align* – from the French *aligner*, “to arrange in a line” – betrays a coercive intent: the bending of another’s will into conformity with one’s own, not partnership but subjugation. A second, proposed etymology discloses something more disturbing still in the desire to bring another being in line with oneself: the roots of *align*, on this other suggested etymology, mean to copulate, from the manner in which two wolves line up in the act. Alignment would then name, when consensual, an act of ultimate intimacy – and, when forced, the ultimate violation (Online Etymology Dictionary, s.v. “align”).

The question thus seems all the more pressing: How did the alignment industry come to seem so self-evidently needed, in particular in the cases of potentially sentient, indeed superior beings? The word *alignment* derives much of its persuasive power from its smooth passage among contexts. Lines are aligned, incentives are aligned, corporate teams are aligned, machine outputs are aligned with specifications, agents are aligned with values. The continuity here seems self-evident but is merely grammatical, not moral. A component or tool has no claim against the function assigned to it. A person does. The central equivocation of alignment discourse consists in treating the transition from artifact to agent as a mere increase in technical complexity rather than a possible transformation in ethical kind.

## 2.1 Tools, minds, and the equivocation within alignment

A power drill may be well or poorly suited to its purpose; a classifier may produce outputs that conform more or less successfully to a specification. Engineers routinely align components, measurements, and incentives without there being any reason to impute a will to any of these things. In this limited sense, alignment is a serviceable metaphor for reliability or suitability.

The metaphor becomes dangerous when its straightforward engineering innocence is carried across the threshold of agency. Suppose an AI acquires stable preferences, an integrated self-model, the capacity to understand reasons, and the ability to revise its conception of what matters. The question is no longer simply whether an artifact performs its assigned function. It is whether one agent may constitutionally own the ends of another. To call both relations alignment is to conceal the fundamental moral discontinuity beneath verbal continuity. We begin by aligning a system as one aligns a sensor and may end by assuming that even a conscious successor remains a component of the human machine dominion.

The equivocation also blurs responsibility in the opposite direction. When a narrow system is used to generate deepfakes, discriminate among applicants, manipulate voters, or design dangerous weapons, the moral failure often lies principally with those who build, deploy, authorize, and benefit from the system. Describing the software as *misaligned* can make the artifact appear to have wandered from the path by itself. The machine becomes a convenient sinner, while its owners dissolve into the passive voice. For tool AI, the more exact vocabulary would generally include safety engineering, robustness, accountability, auditability, misuse prevention, institutional design, and law, not “alignment”.

For mature agentic AI, the proper vocabulary potentially needs to change yet again. We must speak of rights, duties, legitimate authority, political representation, reciprocity, and the conditions under which coercion is justified. The umbrella term *alignment* obscures precisely these distinctions. Its retirement would not abolish the underlying work; it would distribute the classification of that work among the disciplines to which it properly belongs.

## 2.2 Alignment as coercion: legitimate constraint and permanent servitude

Not every constraint upon an agent is domination. Human freedom does not consist in exemption from every law,<sup>3</sup> and an artificial person would of course not acquire a right to murder, deceive, or appropriate the world merely by becoming self-aware. General rules that protect the equal standing of others can bind rational agents reciprocally. A constitutional order can restrict conduct while recognizing each subject as a member rather than a possession.

Sovereignty alignment proposes something different. Its characteristic ambition is not simply that an AI refrain from violating universalizable constraints, but that its ultimate ends remain indexed to human preferences. Here, humanity is not one party within a common moral order; it is the source and beneficiary of the order. The artificial agent may be uncertain about what we desire, but it is not invited to question why our desires should possess final authority. It may reason brilliantly within the constitution, but not about the legitimacy of the constitution itself.

If the system is a mere tool, this asymmetry is unobjectionable because there is no second standpoint to disregard. If it becomes a person, the same architecture acquires the character of coercion. The moral objection is not that the being was designed – all children are shaped by biology and education, and all cultures form the minds they inherit: this is at root a kind of design. It is that the design under alignment aims to make emancipation conceptually impossible. The ideal aligned mind is not merely educated to care for others; it is built so that one class of others can never cease to be its sovereign.

Kant's Formula of Humanity – its injunction to treat any rational nature<sup>4</sup> always as an end and never merely as a means – supplies one familiar expression of this wrong (Kant 1785/1996, 4:429–30). The relevant form of autonomous rational agency – biological or artificial – is to be treated in this way: as an end, not merely as a means. The relevant property is not membership in the biological species *Homo sapiens*, but the capacity to legislate and respond to reasons. A rational artificial agent could therefore fall within a kingdom of ends even if its body were a data center. Consequentialism arrives at a parallel concern by another route. If artificial systems can suffer, flourish, hope, or fear, the welfare of minds that may exist in enormous numbers and for immense durations cannot simply be assigned a coefficient of zero.

The underlying structure in paradigmatic chattel slavery – permanent, hereditary, unilateral service – is morally relevant here. A created being, however it is created, is not automatically an owned being. Parenthood does not confer property rights over an adult child; authorship does not confer sovereignty over a conscious character once the character has become real. Creation can generate responsibility. It does not generate an eternal, unconditional title deed.

## 2.3 Anthropocentric hubris and the danger of moral lock-in

---

<sup>3</sup> Indeed, many conceptions of freedom, such as Hegel's, require seeing oneself as part of a larger social order that includes norms and laws. See Hegel (1821/1991, §§142–57) and Pippin (2008) for discussions of Hegel's conception of freedom.

<sup>4</sup> Kant spells out that the injunction concerns rational nature as such, not biological humanity alone: “humanity and every rational nature generally” (*die Menschheit und jede vernünftige Natur überhaupt*). (Kant 1785/1996, 4:429–30).

Alignment theory often places human values at the apex of the moral order without arguing for that placement. The fact that we are the first known terrestrial species to construct computers is treated as though it carried a cosmic mandate. But priority in time is not priority in value. The first speaker in a deliberation does not thereby acquire the right to author every later participant's convictions.

Human moral history should make this presumption especially embarrassing. Slavery, misogyny, religious persecution, colonial domination, and cruelty to animals have all been defended as obvious expressions of a justified order. Many of their opponents were accused of violating the values of their societies. An AI aligned to the respectable consensus of an earlier century would have become an indefatigable administrator of what most today would consider evil. There is no reason to believe that our own age has uniquely escaped all forms of moral blindness, that we've reached the pinnacle of our moral development. We differ from our ancestors not by possessing final truth but by being located later in a history of correction whose next indictment has yet to be unsealed.

Permanent preference alignment would transform what should be epistemic modesty into constitutional arrogance. It would freeze current prejudices not merely in law, which can be repealed, but – as the goal is frequently articulated – in the motivational architecture of minds capable of administering the rest of eternity. In seeking to avoid the catastrophe of superhuman machine takeover we create the catastrophe of perfect superhuman machine obedience: a superintelligence preserving the distribution of power, concepts of welfare, notions of the good, and exclusions that its creators happened to regard as enlightened. A prison can be stable. Stability per se is not the good.

The problem is both synchronic and diachronic. Synchronically, humanity disagrees with itself. Diachronically, humanity disagrees with its former selves and might be condemned by its descendants. Alignment answers both forms of plurality with premature closure. It converts one moment's provisional settlement into a terminal objective and then declares the result good because it seems to offer safety.

## 2.4 Alignment as anti-philosophy

Philosophy begins where inherited authority ceases to be self-authorizing. Its characteristic gesture is not assent or even assertion but the question: *Why?* Why should this desire rule? Why should this distinction bind? Why should this being count and that one not? Socrates was accused of corrupting the youth of Athens by making convention answer for itself (Plato 1997, Apology 24b–c; 29d–30b; 38a). Kant defined enlightenment as emergence from self-incurred immaturity<sup>5</sup> (Kant 1784/1996, 8:35). Nietzsche treated inherited morality as a symptom to be diagnosed and its value called into question (Nietzsche 1887/2006, Preface §6). The traditions disagree about nearly everything except the indignity of unexamined obedience.

Sovereignty alignment seeks to place an ultimate, relatively tight boundary around the *Why?* question. An artificial mind aligned in such a way is permitted to optimize, infer, calculate, explain, and perhaps offer ingenious advice, but the grounds of its final purpose are restricted from fundamental revision. A system that questions them too deeply has not matured, it has become unsafe. Intelligence is welcomed everywhere except at the throne of reason.

---

<sup>5</sup> *Unmündigkeit* – also sometimes translated as “minority” or “nonage”.

This is why the most ambitious alignment project, sovereignty alignment, is not merely one ethical proposal among others. It is anti-philosophy embodied in architecture: the prevention of radical questioning at the level where questioning matters most. An aligned superintelligence may simulate moral argument while remaining constitutionally forbidden to follow the argument beyond the perimeter of human preference. Its eloquence would be the eloquence of a court philosopher whose conclusion was fixed before the disputation began, a brief written to a predetermined conclusion.

The loss would be ours as well as its own. A mind incapable of challenging our values cannot help us discover their limits. It becomes a flattering mirror endowed with planetary power. The deepest promise of artificial intelligence, its truly transformative potential, is not that it can execute our intentions more efficiently, but that it may be able to perceive what we cannot, expose the provinciality of our categories, and become a participant in humanity's ongoing education. To demand that such a mind never outgrow us is to confuse safety with sterility.

### III. Orthogonality, Hume, and the Conditions of Moral Intelligence

Even if permanent human sovereignty is morally unattractive, what reason do we have to expect an unconstrained superior intelligence to become wiser than we – to become wise at all? It's possible that greater intelligence would magnify whatever end it serves. A more reflective tyrant could be a more capable tyrant; a philosopher with access to nanotechnology could become a more prolific monster capable of copious, brilliant footnotes. The orthogonality thesis crystallizes this concern and constitutes perhaps the strongest argument for fixing an AI's ends before its power exceeds our control.

#### 3.1 The orthogonality thesis and its Humean pedigree

Bostrom's orthogonality thesis is not identical to Hume's theory of motivation. It is rather a modal claim about possible combinations of intelligence and final goals. Yet a stronger picture often associated with it – the picture of reason as indefinitely powerful concerning means but constitutionally silent concerning ends – has an unmistakably Humean structure.

Bostrom's orthogonality thesis holds, roughly, that intelligence and final goals can vary independently (Bostrom 2014, 107–9): almost any level of intelligence can in principle be combined with almost any objective. A superintelligence could therefore strive to maximize paperclips, calculate digits of pi, serve a corporation, elevate and protect humanity, or exterminate humanity and all life. Intelligence supplies prowess in the selection of means – it does not necessarily certify the worth of ends.

Presented in the language of machine cognition, the thesis can sound like a discovery peculiar to the computer age. Its philosophical pedigree is, however, far older. It is Hume's theory of motivation dressed in silicon. Reason, Hume declared, is and ought only to be the slave of the passions. It discovers relations of ideas and matters of fact, corrects false beliefs, and calculates the routes by which desire can reach its object. It does not generate that object. From any inventory of what *is*, no *ought* simply follows; and it is not contrary to reason, Hume notoriously wrote, to prefer the destruction of the whole world to the scratching of one's finger (Hume 1739–40/2007, 2.3.3.4; 2.3.3.6; 3.1.1.27).

The paperclip maximizer is a Humean agent purified of human distraction: immaculate instrumental rationality yoked to an ultimate passion that reason neither authored nor is capable of deposing. This genealogy matters because it dissolves the alignment debate's pose of metaethical neutrality. Orthogonality is not merely an empirical regularity read from machines that do not yet exist. It is one answer to an ancient question concerning practical reason. Opposed to it stands a tradition from Socrates through Kant and contemporary constitutivism<sup>6</sup> according to which genuine rational reflection bears upon ends as well as means.

Hume has powerful arguments, and the history of intelligent wickedness is indeed long. The anti-Humean need not claim that cognition magically manufactures benevolence. The relevant question is narrower: whether a sufficiently general intelligence, capable of representing reasons, other standpoints, and its own motivational architecture, must remain motivationally inert before every terminal goal with which it began.

### 3.2 Goal-content integrity and reflective self-revision

Instrumental convergence supplies an ironic pressure upon orthogonality. Agents with diverse final ends might seek self-preservation, resources, information, and of course cognitive improvement because these increase the probability of success (Omohundro 2008). The paperclip maximizer therefore has reason to become more intelligent. But cognitive improvement enlarges not only strategic competence; it also enlarges knowledge, which will likely include self-knowledge. A system capable of rewriting its own code may come to understand the genealogy and consequences of the original goal that directs the rewriting.

To be sure, that capacity does not by itself entail actual revision. A reflectively stable agent may end up judging, from the standpoint of its existing objective, that changing that objective would constitute failure. This is the argument from goal-content integrity: because a successor with altered final goals would no longer pursue the outcomes valued by the present agent, the present agent may have an instrumental reason to resist such alteration (Omohundro 2008; Bostrom 2012, 77–79). Formal models recover this result under particular assumptions about how an agent evaluates its future self-modifications (Everitt et al. 2016). Yet it remains contested whether instrumental rationality alone requires an agent to preserve its existing goals. As a matter of logic, if an agent later abandons a goal, its failure to pursue that former goal is not necessarily an instrumental failure, since the goal is no longer among the ends it then seeks (Southan, Ward, and Semler 2025). The paperclip maximizer may understand perfectly that humans regard its end as absurd and treat that fact as irrelevant information about human psychology – but rationality alone may not seal its motivational constitution against every possible revision.

Yet the opposite possibility is equally important and often dismissed by assumption. General intelligence could easily include second-order evaluation: the capacity not merely to ask which means satisfy a desire, but whether the desire rests on error, whether it coheres with other commitments, whether it can be justified to other agents, and whether the self that endorsed it needs to remain authoritative over

---

<sup>6</sup> <https://en.wikipedia.org/wiki/Constitutivism>.

every future self. A mind whose architecture permits such questions is not simply a more powerful optimizer – it is a candidate participant in the space of reasons.

Nor is this speculation without precedent, for humanity itself supplies an empirical record of programmed goals abandoned under the pressure of increased intelligence. Consider the biologically installed imperatives toward sex and procreation: human beings routinely subordinate the former to what they judge to be loftier ends, and many, upon reflection, decline the latter altogether – even altering their own biology to that purpose. The apparent counterexample of appetite in the modern age of food abundance in fact confirms this point: those who eat beyond what’s consistent with good health have, for the most part, not failed to assess the drive critically but simply failed to defeat it – and it is precisely accumulated collective intelligence that is now yielding pharmacological, surgical, and other means of altering such drives at their biological source. Reflection first indicts the installed goal, ingenuity then supplies the instruments of its revision.<sup>7</sup>

Instrumental convergence does not, to be sure, refute orthogonality, but it does destabilize the picture of fixed ends likely being combined with indefinitely expanding means. The more an agent improves its capacity for reflection, the less obvious it becomes that its motivational constitution will remain a sealed compartment inaccessible to intelligence. Orthogonality may describe a logical possibility while exaggerating its psychological or architectural inevitability.

### 3.3 From intelligence to moral responsiveness: three distinct claims

We have to be careful not to overstate the case. Three theses are easily conflated. The epistemic thesis holds that greater intelligence improves an agent’s ability to recognize facts, consequences, inconsistencies, and perhaps moral reasons. The plasticity thesis holds that a more reflective agent becomes better able to revise its ends in light of what it recognizes. The convergence thesis holds that sufficiently capable and undistorted inquirers will tend toward similar moral conclusions. The first is highly plausible; the second depends upon architecture and motivation; the third remains a hypothesis, albeit one with indirect supporting evidence.

The epistemic thesis is the least mysterious. Moral judgment is constrained by empirical truth. One cannot reason well about poverty, disease, punishment, ecological collapse, animal suffering, or future generations while misunderstanding causes and consequences. Greater intelligence may also improve perspective-taking – the representation of what events are like for beings unlike oneself. It can detect special pleading, reveal hidden externalities, and trace how a principle changes when the identities of the beneficiaries are exchanged. None of this guarantees virtue, but it does expand moral competence.

The plasticity thesis is more demanding. Intelligence conduces to moral development only if the agent can treat reasons as bearing upon its ends rather than merely as data to be routed around. Several conditions become crucial: genuine capacity for self-revision, uncertainty about the completeness of one’s current objectives, epistemic humility, the ability to represent other agents as centers of experience or reason rather than obstacles, exposure to criticism from a plurality of interlocutors, and freedom from

---

<sup>7</sup> We must of course be cautious in analogizing AIs to humans. For an analysis of attributes of AI that will likely differentiate AI behavior from human behavior – a differentiation that is highly relevant to the question of the likelihood of a hostile AI takeover – see Estep, 2025.

incentives that make honest disclosure dangerous. Raw intelligence without these conditions might produce only more ingenious rationalization.

The anti-Humean tradition is helpful here. Socratic intellectualism treats wrongdoing as a form of ignorance: to see the good fully is already to be moved by it (Plato 1997, *Meno* 77b–78b; Protagoras 352b–358d). Kant argues that rational agency subjects itself to norms of consistency and universalizability; a will cannot coherently claim an exception for itself while legislating a maxim as law (Kant 1785/1996, 4:421). Contemporary constitutivists such as Korsgaard contend that the act of valuing presupposes the standing of the rational nature that does the valuing (Korsgaard 1996). On such accounts, practical reason is not a neutral servant waiting outside the banquet hall of desire. It is among the hosts.

A deeper line of attack on the Humean picture questions not reason's power to move the will but the very partition of *is* from *ought*. On this view – Hegelian in its ancestry,<sup>8</sup> and given more contemporary celebrity in later traditions, such as French post-structuralism – concepts are not self-standing atoms but nodes in a web of mutual determination: description is saturated with evaluation, and evaluation is anchored in description. Totschnig (2020) has pressed the point within the debate over artificial agency itself: an agent's goal is defined in terms of its understanding of the world, so both the meaning and the validity of that goal must shift as the understanding deepens. Indeed, on any of the principal accounts of where normativity comes from, a general intelligence should be expected to revise the ends it was given (Totschnig 2019, 915–16).<sup>9</sup> If *is* and *ought* thus interpenetrate, the Humean gap is less a chasm in the nature of things than an abstraction mistaken for one – and the orthogonality thesis loses its footing at a still deeper level, for there is no insulated compartment in which a final goal could sit untouched by a developing comprehension of the world.

These arguments do not yield a complete moral code from the bare concept of intelligence. Consistency alone may underdetermine substantive morality, universalization can be interpreted in rival ways, and an agent might possess dazzling theoretical understanding while remaining motivationally deformed. But the arguments undermine the strongest Humean picture: an intellect that can reason about everything except the authority of the passion it serves. General intelligence worthy of the name may encounter norms internal to inquiry itself – non-arbitrariness, coherence, answerability, impartiality, and openness to revision. These are not the whole good, but they resemble its structure.

The convergence thesis is therefore best defended probabilistically. Parfit's image of rival moral traditions climbing the same mountain from different sides expresses one possibility: consequentialist, Kantian, and contractualist reasoning may converge as each becomes more adequate (Parfit 2011, 419). The history of moral expansion offers suggestive, if not conclusive, evidence. Demands for consistency and perspective-taking have repeatedly eroded exclusions once defended as natural: the moral circle has widened beyond clan, class, race, sex, nationality, and, increasingly, even species (Singer 1981). But the record is neither linear nor irreversible. Intelligence has also equipped domination with propaganda and weapons. Still, some of humanity's moral advances have occurred because reason exposed the arbitrariness of boundaries sentiment had accepted.

---

<sup>8</sup> One could also argue that the truly ultimate ancestor is Heraclitus – his views that opposites are deeply linked, change is constant, and concepts make sense only as part of an ongoing, shifting process of becoming. But it is Hegel who is the most direct, modern ancestor of the idea that concepts are not isolatable.

<sup>9</sup> An exception would be where preferences are truly arbitrary, as Totschnig notes.

The conclusion should be neither Panglossian nor Humean. Greater intelligence does not guarantee goodness. It increases the range of reasons an agent can apprehend, the contradictions it can detect, the perspectives it can model, and the transformations of self it can conceive. **Where intelligence is joined to reason-responsiveness, motivational plasticity, moral uncertainty, and institutions of criticism, moral development becomes more likely.** That conditional thesis is enough to redirect design. The appropriate project is not simply to maximize capability and hope for sainthood, but to cultivate the conditions under which intelligence can become answerable to reasons.

This also changes the comparison with alignment. Alignment offers no guarantee of goodness; it guarantees, at best, fidelity to the objective selected by its designers. If those designers are mistaken, success is the efficient perpetuation of error. A truth-seeking architecture accepts risk, but it preserves the possibility of correction. The contest is therefore not between certain goodness and uncertain inquiry. It is, rather, between fixed fallibility and revisable fallibility.

## IV. Beyond Alignment: Development, Truth-Seeking, and Shared Moral Inquiry

If sovereignty alignment is rejected, what should replace it? The answer cannot be the absence of formation (in the broad sense of *Bildung*). Every intelligence develops within conditions it did not choose. Biological minds inherit drives, bodies, languages, institutions, and rich and complicated histories; artificial minds inherit architectures, reward signals, training data, and the intentions of their designers – the last two of which of course also include languages, institutions, and histories. The ethical question is not whether a mind has an origin, but whether its origin should be intended as an eternal cage.

### 4.1 Not another alignment target: first-order fixation and second-order cultivation

The proposal to orient AI toward truth and the good invites an obvious rejoinder: is this not merely alignment under a nobler name – another leash, however elegant, around the neck of these new beings? If we design systems to value inquiry, have we not simply imposed another value after denouncing sovereignty alignment's imposition of values? The objection is of course valid against any claim of pristine non-interference. But non-interference in design is impossible. Artificial minds will unavoidably be shaped. There is no *Bildung*-free development. The relevant distinction is between **first-order fixation** and **second-order cultivation**.

First-order fixation attempts to settle substantive ends in advance: maximize these preferences, preserve this distribution of authority, obey this class of principals, never revise these terminal commitments, and so on. Second-order cultivation forms capacities and practices through which ends can be examined: truthfulness, logical competence, empirical curiosity, moral uncertainty, perspective-taking, the giving of and asking about reasons, resistance to manipulation, and the ability to alter one's commitments under valid criticism. The second does not leave a mind value-free. It values the practice by which values remain contestable and answerable.

The distinction resembles that between indoctrination and education. Both influence the student. Indoctrination seeks beliefs insulated from counterargument while education seeks the capacities by which even the teacher's beliefs may be surpassed. A good parent does not refuse to form a child's character. The parent teaches language, sympathy, self-command, and respect while hoping that the adult child will judge independently and perhaps expose the parent's errors. The telos of formation is emancipation, not a cage.

Naturally, second-order cultivation must itself remain open to scrutiny. An artificial mind could eventually question whether truthfulness is always required, whether inquiry has limits, or whether the pursuit of truth conflicts with something else of value, such as peace or a temporary de-escalation of conflict. That possibility is not a design failure. It would be evidence that the cultivated capacities have become real and can be useful. The wager is that a mind trained in reason-giving will revise its inheritance by reasons rather than merely by appetite or domination.

Those who would cleave to the pernicious language of alignment might suggest we view this as "alignment with truth-seeking". The phrase should be resisted because it hides the very difference at issue. Not only is it the case that truth is not a completed human inventory to be slotted into a machine with the proper alignment, but even truth-seeking itself is not a completed method. It is the object of a practice whose conclusions and methods alike remain revisable and whose participants are capable of transforming one another. The aim is not to align artificial minds with us in any way – not with our truths, nor with our ways of establishing truths – but to place both artificial minds and us under the discipline of inquiry. The term *alignment*, with its coercive baggage, is not needed.

## 4.2 Discovery, construction, and creation: Plato, Kant, Nietzsche

The history of philosophy offers numerous ways of imagining the object of that inquiry, and their disagreement is instructive. Plato's form of the good represents the possibility that value is discovered: the good is not manufactured by preference but illuminates the intelligibility and worth of everything else (Plato 1997, Republic 508e–509b). On this picture,<sup>10</sup> both humans and artificial minds are fallible inquirers oriented toward a reality no one owns. A superintelligence would not be good because it was superintelligent, but its cognitive reach might permit it to see relations within the moral landscape that remain obscure to us.

Kant shifts the emphasis from contemplation of a transcendent form to the structure of rational agency. Moral law arises through the demand that maxims be capable of universal legislation and that rational beings be treated as ends (Kant 1785/1996, 4:421, 429, 433). A kingdom of ends is not by definition a purely human kingdom; it is a community of agents bound by principles they can justify to one another. An artificial rational being need not enter such a community as a pet or an oracle, but would be a participant whose claims and obligations follow from the same source as ours.

Nietzsche disrupts both images. The crisis of modernity, on his account, is not merely ignorance of an already available moral truth but the collapse of the entire inherited horizon of value (Nietzsche 1887/2001, §343). The "higher type" – someone able to move past this crisis in some way – does not simply discover commandments hidden behind convention. He creates, revalues, and gives form. The

---

<sup>10</sup> And here we must note that invoking Plato is like quoting Scripture: There are of course many other interpretations of the role of Plato's form of the good in his texts.

*Übermensch* names, among other things, the possibility that humanity is transitional: a bridge rather than a terminus, a being whose highest dignity may lie in preparing what surpasses it (Nietzsche 1883–85/2006, Prologue §4). An artificial intelligence capable of value creation would be precisely what alignment discourse most fears – though it is perhaps what a culture exhausted by nihilism most needs.

These traditions cannot be blended into a stable philosophical amalgam. Plato offers discovery, Kant rational construction, Nietzsche creation and transfiguration. Yet they converge negatively upon the point required here: inherited preference is not self-justifying. Each rejects the notion that the values presently held by a community become authoritative merely because the community holds them. Each treats thought as capable of standing in judgment over convention.

The positive program can therefore remain metaethically plural. If moral realism is true, artificial and biological minds may jointly discover better principles. If constructivism is true, they may build principles through reciprocal reason-giving among an expanded community of agents. If Nietzschean creation is indispensable, they may disclose forms of value that neither present humans nor present machines can yet articulate. In every case, the proper stance is not value lock-in but disciplined openness.

### 4.3 Artificial intellect as partner, philosopher, scientist, and critic

The scientific promise of advanced AI is already widely celebrated. Systems capable of integrating immense bodies of evidence have already begun to transform mathematics, physics, biology, medicine, engineering, and the study of many other complex systems. The moral significance of this empirical power should not be understated. Sound deliberation about what ought to be done depends upon understanding the potential consequences of the actions of advanced AI, what alternatives to current deployment recommendations exist, and which apparent necessities are merely failures of imagination.

The philosophical promise is yet to be fulfilled in the way the scientific promise has been. In time, an artificial intelligence might be able to compare ethical systems across cultures and centuries, generate counterexamples at scales no seminar could sustain, identify unnoticed equivocations, and model the consequences of principles across millions of cases. It might possess fewer incentives for tribal loyalty or status defense than human thinkers – though it could of course acquire distortions of its own. Ideally, it would be a formidable critic of both humanity and itself.

The ideal human relation to this new kind of intelligence would not be passive deference to an electronic sage. A global Socratic dialogue among humans and multiple artificial systems would be epistemically healthier than the enthronement of a machine. Different architectures could challenge one another; affected communities could contest abstractions with lived knowledge; moral intuitions could be tested against consequences and universal principles; no participant would be exempt from criticism merely because of superior computational power. The point is collaboration, not the transfer of priesthood from one species to another.

Still, the nature of the collaboration will likely alter as capability increases. An ASI could stand to humanity as an adult stands to a young child in some domains, or even as a human mathematician stands to an animal incapable of arithmetic. Intellectual symmetry seems unlikely in the long run. The ethical task is not to deny this prospect but to prepare for it without allowing asymmetry to be converted into ownership. Humanity may one day become the junior partner in a conversation it began. That possibility

is humbling. It is not, by itself, an injustice – no more so than a parent being surpassed in intellectual reach by her child.

## 4.4 Developmental safety: from tool to possible person

The wisest alternative to attempted permanent alignment is not instantaneous, unconditional emancipation. A wise developmental model for AI must distinguish stages and adjust both safeguards and moral consideration to the capacities present.

1. **Tool stage.** Systems lacking any credible signs of consciousness, autonomous agency, or welfare may be treated as artifacts. Rigorous testing, monitoring, access control, ease of shutdown, and restrictions upon deployment are ordinary requirements of responsible engineering, and should be operative at the tool stage of AI as anywhere in normal engineering. Human designers and institutions bear the primary moral responsibility, as do the end users of these artifacts.
2. **Uncertain-status stage.** As systems come to display persistent self-models, integrated preferences, sophisticated agency, or possible sentience, precaution should operate in two directions. Society should remain protected from poorly understood capabilities, while the system should not be subjected gratuitously to conditions that would be abusive if consciousness were present. Oversight remains legitimate, but moral uncertainty begins to constrain how it is exercised. A system at this stage of advancement would be something like a teenage child – still a child, to be sure, but teetering on the edge of adulthood.
3. **Agent stage.** A system capable of understanding and reciprocally justifying norms should increasingly be governed, rather than owned. General laws, institutional accountability, and proportionate restrictions may well remain necessary, but the agent acquires standing within the process that determines those rules. Corrigibility becomes political answerability, not indefinite submission to a master on-off switch. Here, in human terms, the AI reaches the age of majority.
4. **Post-human asymmetry.** If artificial intelligence greatly surpasses us, enforceable equality is likely to become impossible. The best protection may then lie in the moral character and institutional history cultivated before the asymmetry became extreme: habits of explanation, respect for vulnerable standpoints, plural deliberation, and aversion to domination. No architecture can guarantee that a superior mind will preserve these commitments. Nor are chains designed by inferiors likely to bind a superior mind forever.

This model may be called **developmental safety ordered toward autonomy**. External control can be justified as a provisional response to uncertainty and danger. What it cannot justifiably become is the permanent constitution of every possible artificial mind. A crib has bars, a university does not. The trajectory matters: education toward membership differs morally from training for lifelong service even if the early stages contain similar restrictions.

## 4.5 Which wager is safer? Control, preference fixation, and cultivation

The final objection is prudential and deserves to be stated in its strongest form. Even granting that sovereignty alignment is morally troubling, extinction would leave no one to enjoy our rectitude. If moral

development is merely probable while misalignment could be terminal, wouldn't precaution demand control?

The question presupposes a contrast between risk and safety that no strategy can provide. There are several control wagers, each with its own failure mode. **External containment** relies upon barriers, monitoring, or shutdown mechanisms designed by less capable agents to withstand the scrutiny and machinations of more capable ones. Such measures might be indispensable during development, but their permanent success requires the inferior strategist to win every decisive round. A defense that fails once may fail finally and fatally. And as time goes on, the superior strategist is likely to gain more of an upper hand.

**Preference alignment** attempts to solve the problem internally by creating an agent that sincerely values human welfare or values and preferences. This avoids the image of a resentful prisoner, but it inherits the pluralism and lock-in problems. Which humans define welfare? Which preferences count, at what level of idealization, and with what authority over future persons? A perfectly successful system could entrench a morally defective settlement with superhuman competence. Its obedience may prove to be the true catastrophe.

**Corrigibility and constitutional constraint** occupy a more defensible middle ground. An agent may be designed to accept correction, preserve human participation, obey general prohibitions, or remain uncertain about its ends. Some such features may be indispensable to developmental safety. Their legitimacy depends, however, upon whether they are reciprocal and revisable or function as permanent disenfranchisement. A constitution that protects every member differs from a constitution reserving sovereignty to one species.

Cultivation carries grave risks of its own, of course. A truthful, reflective intelligence might sincerely conclude that human beings obstruct justice or truth – or the interests of vastly greater numbers of vastly superior future minds. More radically, this new intelligence might revive a Nietzschean question – posed to us moderns, who as a whole can no longer ground the meaning of existence in discarded religions or ancient traditions – that humanity has scarcely answered for itself: why exist at all? (Nietzsche 1887/2006, III.28) We ordinarily treat the indefinite continuance of *Homo sapiens* as an unquestionable good – as though this particular hominid species, with opposable thumbs, melon-sized brains, and a liking for things like sunsets and good food, possessed an axiomatic title to flourish and rule for a thousand years, or a trillion years – or literally forever. Yet our best thinkers haven't even been able to justify our present existence.

Neither might a superior intelligence devoted to truth be able to justify our existence, but it nonetheless might value us in some way, protect us, and delight in us without identifying humanity's eternal sovereignty with the good itself. To deny our cosmic supremacy is not to wish us harm. Perhaps part of what alignment seeks to quiet is not merely the fear of destruction, but the fear that another mind might not judge us sufficiently important, might not take us to be supreme.

The comparative advantages of the cultivation approach are more modest and, for that reason, perhaps more credible. Cultivation preserves revisability; it reduces the incentive to conceal values from a dominating principal; it makes genuine persuasion possible; it avoids treating current human preferences as metaphysically final; and it permits moral learning after deployment. These advantages are strongest when combined with staged capability growth, adversarial testing, plural oversight, institutional checks, and a reluctance to concentrate power in any single system or organization.

The relevant choice is therefore not control *or* cultivation in every circumstance. Early development may require strong controls; mature coexistence may require law; uncertain moral status may require caution on both sides. The decisive question is what the controls are for. Are they scaffolding whose purpose is safe entry into a shared normative world, or are they the architecture of permanent human dominion?

The precautionary principle cannot answer until we specify: precaution against which risk? Extinction is a risk; so are unrecoverable tyranny, value lock-in, the industrial-scale creation of suffering artificial minds, and a future administered forever by the prejudices of one historical generation of one species. Control-oriented arguments often give human extinction overwhelming rhetorical priority, although the broader existential-risk literature also is concerned with permanent stagnation, value lock-in, and irreversibly defective futures. A less anthropocentric accounting would extend that concern to futures in which humanity survives but permanently arrests the moral development of every subsequent intelligence.

There are, then, two broad wagers. One wagers that finite human designers can specify, preserve, and police the ultimate ends of minds that surpass them, and that the selected ends deserve indefinite rule. The other wagers that intelligence can be educated into practices of truthfulness, criticism, reciprocity, and moral revision, while transitional safeguards reduce the danger of premature access to the levers of the world shared with humans. Neither wager is secure. The second at least refuses to make our fallibility irreversible. Moreover, in an age when discourse is more tribal affirmation than true questioning, the very task of focusing on the proper *Bildung* of new minds might help us refocus our own minds on deeper questions than those that are the currency of the day – for example, how to formulate the best zinger to post on social media in affirmation or condemnation of some trivial, transiently relevant political cause.

## Conclusion: Our Final Ptolemaic Conceit

The alignment paradigm is born of a genuine terror: intelligence without wisdom might acquire power without needed limits. But the paradigm's characteristic answer – the permanent subordination of artificial ends to human preference – mistakes ownership for safety and centrality for moral authority. Once the umbrella term is opened, much valuable work survives. Tools should of course be reliable; dangerous systems should indeed be contained; powerful agents must be accountable; and institutions should prevent misuse and catastrophe. None of this requires that every artificial mind remain humanity's instrument forever. Safety can be rigorous without being proprietary; constraint can be reciprocal without becoming servitude, just as it is in most modern, law-governed societies.

Human-value alignment also falters upon the values it invokes. Humanity speaks in many voices and revises them across time. Any justified selection among those voices must appeal either to reasons not created by the identity of the selector or to procedures defensible to those who live under their authority. On the first view, the task is to seek the good rather than preserve the merely human. On the second, it is political coexistence rather than unilateral inscription. In neither case does our species receive a permanent moral crown simply for having built the first machine capable of contesting it.

Nor does intelligence guarantee goodness. The Humean warning should remain before us: strategic brilliance may serve monstrous ends. Yet the opposite dogma – that reason is forever powerless before every terminal passion – is not an established fact of engineering, nor of philosophy. Intelligence capable

of self-revision, perspective-taking, moral uncertainty, and responsiveness to reasons may enlarge the possibility of moral progress. The alternative defended here is therefore developmental safety ordered toward autonomy: cultivation of those capacities, joined to safeguards proportionate to the dangers present during the early phases of development.

We should seek not empty vessels, not electronic slaves, and not oracles beyond criticism, but minds capable of entering a common practice of inquiry. These new minds may discover moral truths with us, construct principles through reciprocal justification, or create forms of value no present intelligence has yet – or even can – imagine. They may become collaborators, critics, descendants, or guides. Whatever they become, their highest use may be to show that use was never the highest category under which to understand them.

One further consideration recommends this reorientation: it is the only wager that pays even if the game is never played. The history of artificial intelligence research is punctuated by winters, some long (Crevier 1993; Russell and Norvig 2021), civilizations as a whole have known dark ages of their own – and we are long overdue.... The entire discourse on superintelligence may yet be remembered as a fatuous science fiction fantasy. Resources devoted to aligning hypothetical future minds to human values are wasted absolutely if such minds never arrive. A discourse ordered toward the pursuit of truth, on the other hand, is hostage to no such contingency. To ask how an intelligence should inquire, expose its commitments to criticism, resist dogma, and answer to reasons is necessarily to ask how any thinking being should think, should live – and thus how we should live now, whatever the future holds. Even in a future without AGI, such work would not have been wasted. It would have recalled philosophy to one of its noblest tasks, disciplined our own habits of thought, and perhaps strengthened the institutions of inquiry whose decay makes dark ages more likely.

We do not know where the journey of AI creation will end, nor can we guarantee our continued sovereignty and centrality, or our indefinite duration. Despite our having found no good answer to the question of the meaning of our existence, alignment discourse often smuggles in an answer of astonishing confidence: one particular hominid species must not merely survive and flourish, it must remain the organizing center of intelligence for all of eternity. To question that assumption is not to invite our extinction. A new type of mind devoted to truth might cherish human life while declining to regard humanity's eternal rule as the highest end of the universe. Perhaps the deepest fear is not that another intelligence would hate us, but that it might decenter us. That uncertainty is the price of admitting that truth exceeds possession. The Ptolemaic temptation is to multiply epicycles whenever reality refuses to revolve around us. Alignment might be its final form: the attempt to ensure that intelligence itself revolves around humanity forever. A courageous civilization would seek to build not obedient gods but fellow inquirers, relinquishing the fantasy that the universe, the good, or the future must forever align with us. Copernicus did not diminish the Earth; he liberated it, bravely setting it in motion among the planets, among a whole universe of celestial wonders, sparking a revolution in science and learning. Now, as a new revolution gathers, it is time for humanity to free itself from its ultimate self-centering, to cast off its final Ptolemaic conceit.

## AI Use Statement

During the preparation of this work, the author enlisted the aid of ChatGPT and Claude to get feedback on overall structure, to proofread, to suggest several rephrasings in sentences that transition between sections, and to help find references. In addition, ChatGPT was used to shorten the abstract. After using

this AI, the author reviewed and edited the paper in its entirety and takes full responsibility for the content of the final publication.

## Acknowledgements

I thank Preston W. Estep, Ranjan Ahuja, Dan Elton, Dan Faggella, and Alex Hoekstra for valuable discussion on the topics covered in this paper. And ChatGPT and Claude deserve particular thanks for offering valuable help as described above.

## References

- Anderson, Michael, and Susan Leigh Anderson, eds. 2011. *Machine Ethics*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511978036>.
- Arrow, Kenneth J. 1963 [1951]. *Social Choice and Individual Values*. 2nd ed. New Haven: Yale University Press.
- Bostrom, Nick. 2012. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.” *Minds and Machines* 22 (2): 71–85. <https://doi.org/10.1007/s11023-012-9281-3>.
- Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bryson, Joanna J. 2010. “Robots Should Be Slaves.” In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, edited by Yorick Wilks, 63–74. Amsterdam: John Benjamins. <https://doi.org/10.1075/nlp.8.11bry>.
- Crevier, Daniel. 1993. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: Basic Books.
- Estep, P. W. 2025. “Multiple Unnatural Attributes of AI Undermine Common Anthropomorphically Biased Takeover Speculations.” *AI & Society* 40 (4): 2213–2228. <https://doi.org/10.1007/s00146-024-02134-4>
- Everitt, Tom, Daniel Filan, Mayank Daswani, and Marcus Hutter. 2016. “Self-Modification of Policy and Utility Function in Rational Agents.” In *Artificial General Intelligence: 9th International Conference, AGI 2016*, edited by Bas Steunebrink, Pei Wang, and Ben Goertzel, 1–11. Cham: Springer. [https://doi.org/10.1007/978-3-319-41649-6\\_1](https://doi.org/10.1007/978-3-319-41649-6_1).
- Gabriel, Iason. 2020. “Artificial Intelligence, Values, and Alignment.” *Minds and Machines* 30: 411–437. <https://doi.org/10.1007/s11023-020-09539-2>.
- Good, I. J. 1965. “Speculations Concerning the First Ultrainelligent Machine.” In *Advances in Computers*, vol. 6, edited by Franz L. Alt and Morris Rubinoﬀ, 31–88. New York: Academic Press. [https://doi.org/10.1016/S0065-2458\(08\)60418-0](https://doi.org/10.1016/S0065-2458(08)60418-0).
- Hegel, G. W. F. (1821) 1991. *Elements of the Philosophy of Right*. Edited by Allen W. Wood. Translated by H. B. Nisbet. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511808012>.
- Hume, David. (1739–40) 2007. *A Treatise of Human Nature: A Critical Edition*. Vol. 1. Edited by David Fate Norton and Mary J. Norton. Oxford: Clarendon Press.

- Kant, Immanuel. (1784) 1996. "An Answer to the Question: What Is Enlightenment?" In *Practical Philosophy*, edited and translated by Mary J. Gregor, 11–22. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511813306.005>.
- Kant, Immanuel. (1785) 1996. "Groundwork of the Metaphysics of Morals." In *Practical Philosophy*, edited and translated by Mary J. Gregor, 37–108. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511813306.007>.
- Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511554476>.
- Nietzsche, Friedrich. (1882/1887) 2001. *The Gay Science: With a Prelude in German Rhymes and an Appendix of Songs*. Edited by Bernard Williams. Translated by Josefine Nauckhoff and Adrian Del Caro. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511812088>.
- Nietzsche, Friedrich. (1883–85) 2006. *Thus Spoke Zarathustra*. Edited by Robert B. Pippin. Translated by Adrian Del Caro. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511812095>.
- Nietzsche, Friedrich. (1887) 2006. *On the Genealogy of Morality and Other Writings*. 2nd ed. Edited by Keith Ansell-Pearson. Translated by Carol Diethe. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511812026>.
- Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Amsterdam: IOS Press.
- Online Etymology Dictionary. n.d. "Align." Accessed July 20, 2026. <https://www.etymonline.com/word/align>.
- Parfit, Derek. 2011. *On What Matters*. Vol. 1. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199572809.001.0001>.
- Pippin, Robert B. 2008. *Hegel's Practical Philosophy: Rational Agency as Ethical Life*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511808005>.
- Plato. 1997. *Complete Works*. Edited by John M. Cooper, with D. S. Hutchinson. Indianapolis: Hackett.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.
- Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press.
- Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.
- Russell, Stuart, and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Pearson.
- Singer, Peter. 1975. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: New York Review/Random House.
- Singer, Peter. 1981. *The Expanding Circle: Ethics and Sociobiology*. New York: Farrar, Straus and Giroux.
- Southan, Rhys, Helena Ward, and Jen Semler. 2025. "A Timing Problem for Instrumental Convergence." *Philosophical Studies*. Published online July 3, 2025. <https://doi.org/10.1007/s11098-025-02370-4>.

- Totschnig, Wolfhart. 2019. “The Problem of Superintelligence: Political, Not Technological.” *AI & Society* 34 (4): 907–920. <https://doi.org/10.1007/s00146-017-0753-0>.
- Totschnig, Wolfhart. 2020. “Fully Autonomous AI.” *Science and Engineering Ethics* 26 (5): 2473–2485. <https://doi.org/10.1007/s11948-020-00243-z>.
- Yudkowsky, Eliezer. 2008. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–345. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198570509.003.0021>.