



AI: Standing at the Crossroads

Brendan McCord

About the Author

Brendan McCord is founder and Chair of Cosmos Institute, a pioneering academy for philosopher-builders, which launched in 2024 with the creation of the Human-Centered AI Lab at the University of Oxford. He co-delivers an AI and philosophy graduate seminar at Oxford, where he was also a Visiting Fellow in philosophy at St Catherine's College. Previously, he founded two AI startups, served as President of an AI lab within an investment holding company, and as President of Technology at the fastest-growing company in its industry's history. In the public sector, he founded the first applied AI organisation for the US Department of Defense, and has launched two record-setting open source AI projects. He holds degrees from MIT and Harvard Business School.

This paper was edited by Benedict Rogers, Joshua McNicholas, and Anna Webb.

Standing at the Crossroads

Brendan McCord

For as long as humans have walked the planet, we have been the most impressive, intelligent beings. Now, for the first time, we face the prospect of artificial minds potentially surpassing our capabilities—a predicament no previous civilisation has encountered.

The rise of artificial intelligence (“AI”) has thrust us into a “civilisational moment”,¹ challenging our conventional frameworks for understanding the world. This crisis reveals itself subtly at first—a growing unease as we recognise that our usual technical, business, and policy tools fall short.

Technologists at top AI labs are at the forefront of this realisation, grappling with dilemmas that extend far beyond code. Jack Clark, Anthropic co-founder, and Cosmos Founding Fellow, illustrates this new awareness: “The more hardline you tend to be about AI safety... the more comfortable you get with advocating for extremely authoritarian policy interventions... or even kinetic actions (bombing data centres).”² Clark’s observation captures the latest form of a very old tension, that between security and freedom. But what drives the complexity of our situation beyond traditional frameworks is that the unknowns around AI are pushing AI risk speculation in multiple, often contradictory directions.

While periods of stability often relegate fundamental questions to the background, pivotal moments in history bring them to the fore. Moments like the American Founding or French Revolution prompted deep philosophical reflection on the nature of government and the rights of men. Scientific revolutions—Copernicus dislodging us from the centre of the universe, Darwin re-situating us among the animals, or Einstein challenging the limits of our senses and intuition—each forced a profound reckoning with our understanding of the world and our place in it. Today, in the age of Turing, we are compelled to revisit these concerns: what does it mean to be human in the midst of artificial intelligence? What does it mean to live a good human life when artificial minds can do what we do, and in some cases much better?

These inquiries strain our civilisational inheritance. We cannot turn to Herodotus, Livy, Voltaire, or even Smith, Mill, or Hayek for direct guidance on navigating intelligence *in silico*. Some argue that this AI-induced civilisational moment demands a radical reorientation, a blank slate approach. Others advocate for improvised philosophies like

a tragic “Prometheanism” or a hopeful “Transhumanism”. We propose a different path.

Rather than abandoning our intellectual traditions or embracing these new philosophies, we assert the urgent need to return to the fundamental questions that have shaped Western civilisation. Paradoxically, the philosophers of bygone eras, whose work predated AI by centuries or millennia and who never contemplated “the alignment problem” or calculated a “P(doom)”, may prove our most valuable guides in this uncharted territory.³

Currently, two primary philosophies dominate the AI discourse. The first is existential pessimism—the view that AI is a dire threat to humanity that requires a “pause” in development.⁴ The opposing school, accelerationism, embraces the power of markets and optimism, but often views technology as the sole end of human life or looks forward to a moment when humans are superseded altogether.⁵

While both philosophies offer valuable insights, they are ultimately insufficient. They often neglect the most fundamental questions about human flourishing and lack a rich enough understanding of the human goods we wish to preserve and defend, which extend far beyond the avoidance of death or the maximisation of technological advancement.

Autonomy—the freedom to think and act independently—is central to human flourishing. The grand ambition of progress has been to increase this freedom, allowing individuals to educate themselves, live according to their own choices, and realise their aspirations. This ideal depends on both the expansion of freedom from undue external control and the internal freedom to develop our capacities fully.

As we increasingly rely on technology, particularly through “cognitive offloading” to AI-driven systems, our autonomy faces profound new challenges that demand philosophical and technological responses. Externally, new technology governance models, driven by fears about AI safety, calls for a particular vision of fairness, or the desire to centralise power, can subtly erode individual freedom of action. Internally, AI’s promise of convenience tempts us to surrender the hard work of thinking and deciding for ourselves, potentially undermining the self-mastery, rational deliberation, and courage necessary to align our actions with our highest capacities—traits that Aristotle saw as essential to genuine autonomy.⁶

Autonomy is not only a human good constitutive of our happiness and flourishing, but also a key source of the West’s continued dynamism. To safeguard autonomy in the age of AI, we must examine the claims of leading schools of thought, analyse their fundamental assumptions, and extract their valuable insights, while complementing them with a renewed focus on human autonomy and flourishing.

AI has catalysed the West's already simmering need to come to terms with its identity. Looking past its triumph after the Cold War, the West has struggled to renew its purpose and meaning. Now, AI and the digital frontier have accelerated the inevitable reckoning. Charting a path forward on AI must align with the West's foundational principle of autonomy as the cornerstone of its dynamic social order.

The West is a Dynamic Order

To appreciate the risks of AI and, more precisely, the dangers posed by *our responses to AI*, we need to understand a basic fact about Western Civilisation. The West is not a monoculture. It is not one thing, one set of laws, one religion, one book, one scientific method, or one belief system.

On the contrary, the West is a dynamic, complex, emergent order. The tension between its essential elements drives its ever-changing kaleidoscope of configurations.

The first root of the West lies in the archaic worlds of Greece and Mesopotamia, with the epics and ethos of the Bronze Age. From here, we inherit the heroic archetype and its celebration of adventure, friendship, and glory as the core of the best life. But we also inherit two critical responses to these ideals. One comes from Greek philosophy, especially the Socratic and, later, the Aristotelian school, which puts reason and science at the centre of civilisation. The second comes from the ancient Israelites and Biblical revelation, which replaces the hero with the patriarch and celebrates the family and obedient love as the key elements of human flourishing.

For millennia, the dynamics of Western history have transformed these basic elements into new and vital configurations. Whether in Rome or Antioch, Alexandria, Paris, Oxford, or Vienna, each capital of Western culture has formed its own local synthesis, giving rise to a new vision, which is often later replaced, sometimes violently, by an even newer synthesis.

The same is true for modern Western liberalism. Western liberalism, focusing on the individual, works for us because it taps into the primary overlapping agreement between our three pillar traditions. All agree on the importance of the individual, i.e., that the subversion of individual flourishing (as they each understand it) does violence to the natural possibilities of human excellence.

All three traditions agree that the power to decide the question of the ultimate good must finally rest with the individual and not with the community, because the latter cannot compel the individual in the one thing that matters—be it heroic striving, free inquiry, or obedient love. Although the political order may be *supportive* (or even *subversive*) of these activities, ultimately, Western liberalism converged on the idea that the individual ought to be sovereign in attempting to realise these ideals.⁷

What makes the Western vision for civilisation daring is precisely this: it is willing to believe that individuals should be left to form bonds among themselves to support individual excellence and that *these bonds* are the functional basis for governance, not a top-down, totalising, and prescriptive order. In fact, governance can become more secure if it invests in individual autonomy and promotes an ideal of individual responsibility—including the freedom to explore the versions of the best life recommended by the West’s competing traditions.

This bet on Western dynamism and its focus on the individual has paid enormous dividends in America. If the measure of a civilisation is the extraordinary character and range of the excellence of its individuals, then the pantheon of American greatness supplies ample proof of its success.⁸

Part of what has made America so remarkable is the vast benefits reaped from its commitment to scientific and technological progress. No other nation in history has built up so much wealth so quickly and through such great technological prowess.

But the digital frontier, especially artificial intelligence, adds new tension to Western dynamism. The storm of opinion surrounding AI is bringing forth ideas that are very much at odds with the core tenets of American liberalism and its synthesis of Western ideals.

Notably, the dominant opinions on AI contain fragments or patterns of ideas that the West rejected in its convergence on the individual. And those who dominate the conversation about AI subsume individual autonomy within frameworks of safety-oriented world governance or reductionist transhumanism. To put this differently, AI is forcing the West to come to terms with its “shadow”—forms of belief that it once struggled with and continue to haunt its primary systems. Interrogating the AI opinion landscape will help us see why the extreme positions are dubious on their own terms and why they threaten to pull us into worldviews incompatible with Western vitality.

AI innovation is advancing rapidly and forcing urgent decisions about policy and norms, a problem exacerbated by our lack of historical precedent. This pace and uncertainty tempts many to trade Western beliefs for the promise of greater security or technological transcendence. However, these trades are not only premised on ill-founded notions; they involve rejecting principles that have allowed Western dynamism to renew itself to the great benefit of the individual.

The AI Philosophy Landscape

The two pictures of AI dominating our epistemic commons are oriented by extreme pessimism and optimism. We will refer to these as “existential pessimism” and “accelerationism”.

Existential Pessimism

Existential pessimism has emerged as today’s intellectual incumbent. Its various schools and leading thinkers are by far the most potent sources of influence on the broader implications of AI for human life. Together, these players have sold more books, raised more money, created more institutes and programmes, and built more policy coalitions than any other alternative.⁹ In terms of its reach and impact, this school of thought may be the most practically adopted intellectual movement since Marxism.

At its core, existential pessimism worries about the future of our planet, its life forms, and, of course, us humans, especially the things we humans do to throw our planetary balance into jeopardy. According to this school’s standard narrative, human activity has most endangered our stability, with the main culprits being the pollution generated by industrial technology and the threats posed by biological, chemical, and nuclear weapons. Artificial intelligence is the latest addition to the technologies of existential risk, which, in its possible “god-like” or “superintelligent” form, may overtake us as the dominant entity on planet Earth.

Within the broader ideology of existential pessimism, we can identify three main, now deeply intertwined, sub-schools: rationalism, long-termism, and effective altruism.

A. Rationalism

Rationalism arose with two major blogs popular in the 2000s: the so-called “Sequences” of Eliezer Yudkowsky and Scott Alexander’s “Slate Star Codex”. Rationalism here is ostensibly about being rational, which, on its face, is consistent with at least one major strand of Western dynamism, the so-called Socratic tradition. But instead of pursuing the various modes of reason humans have developed, it weds “rationalism” to a specific, even myopic notion of being reasonable, centred on Bayesian updating with a value function over outcomes.¹⁰ This narrowing of rationality would have been alien to Plato, Aristotle, and many modern thinkers such as Mill and Kant.¹¹

Rationalists have also discovered that it is not easy to escape the effects of mood, desire, and passion. Despite its worries about bias and being “[less wrong](#)”,¹² the formal character of “Bayesianism” has, ironically, proven fruitful soil for the extremes of hope

and despair. In contrast to thinkers like Plato, Hume, Rousseau, or Nietzsche, who are self-conscious in thinking through the relation of reason to its various roots in human psychology, Bayesian formalism does not do enough to question the premises of thinking itself, including the charms of applying abstract, quantitative approaches to problem-spaces that do not seem especially suited to this sort of reasoning.

In the case of Yudkowsky, his roughly Bayesian approach failed to filter out some of the driving influence of his personality. This is evident in his extreme positions on AI risk, which often seem to stem more from personal fears than dispassionate analysis. Consider the *Time Magazine* article “The Only Way to Deal with the Threat From AI? Shut It Down”.¹³ Here, Yudkowsky refers to an open letter set up by Max Tegmark’s Future of Life Institute calling for a six-month moratorium on “Giant AI Experiments”.¹⁴ For Yudkowsky, the letter is a good start, but it is not sober enough. What it misses is that “smarter than human” AI is extremely hard to detect. And given the likely negative outcome—“under anything remotely like the current circumstances... literally everyone on Earth will die”—Yudkowsky takes the subtlety of identifying and subordinating the emergence of superintelligence with deadly seriousness, even to the extent that appears paranoid or hysterical.¹⁵

Yudkowsky’s approach to AI risk is further complicated by his concerns about unintended consequences. He worries that his current statements about AI risk and potential apocalyptic outcomes might inadvertently provide information to a future malicious AI. Like some sort of Golem that learns from the nightmares of its maker, Yudkowsky believes that what we say now about the worst-case AI scenarios will somehow inadvertently script our AI fate. This is why he was adamant about the need for an outright ban on AI research—for little open discussion of AI risk and what we can do about it—and why he seemed so supportive of total and drastic measures, such as airstrikes against data centres as part of our defensive response.¹⁶

These extra layers of self-conscious concern contribute to the perception of seemingly unconstrained worry, which, to many, appears at odds with the rational approach Yudkowsky advocates. These impressions have hurt Yudkowsky’s status, especially because he is most identified with the so-called “Rationalist” movement, or “Less-Wrong” movement, which prides itself on Bayesian sobriety and evidence-based reasoning. The fallout has been significant, with individuals who have relied on Yudkowsky’s assessments of AI risk, including influencers in the effective altruism movement, facing scrutiny by association.¹⁷

B. Long-termism

It is hard to treat Nicholas Bostrom separately from Yudkowsky; both work on similar problems and have often been doing so together since the early 2000s. While both Bostrom and Yudkowsky are now famous for their views on AI, Bostrom’s intellectual journey began with Transhumanism, which led to his first major foothold in existential

risk. This trajectory differs from Yudkowsky's, who has focused more narrowly on artificial intelligence throughout his career.

Bostrom's wider scope is evident in his work on various threats, from natural disasters to advanced technologies. His long-termism worries about the existential risks posed to civilisation by whatever might threaten it.¹⁸ In this sense, his work fits into a tradition of history and political thought focusing on regime cycles and patterns of ascent and decline. Classical thinkers like Plato and Lucretius address the problem of civilisational collapse and even include technology as a risk to political stability.¹⁹

Bostrom's sobering presentations of the problems posed by speculative technologies make him stand out in this tradition. For Bostrom, civilisation should worry about the broad category of existential threats, many of which are natural, like asteroids or viruses. As civilisations grow more complex and powerful, however, they must contend with the effects of their own technological advancements. The multivalent nature of technology means that tools capable of mastering nature may also pose risks to the very cultures that created them.

Bostrom illustrates technological society's vulnerability to the risks of innovation through an analogy of repeatedly drawing balls from an urn. Most balls drawn are either beneficial ("white") or have negligible negative effects ("grey"), representing technologies that improve or at least do not harm society. However, as technological progress accelerates and becomes more complex, the probability of drawing a "black ball" increases. Here, a black ball technology is any system or device that might be powerful enough to degrade or eliminate civilisation—something like a massively infectious synthetic virus or a volatile form of nuclear fusion.²⁰

For Bostrom, artificial general intelligence or "superintelligence" is perhaps the ultimate "black ball" technology. A key feature of superintelligence is that it can build the next version of itself in ways that eventually exceed human comprehension. Or, to put it differently, a superintelligent system will not only outperform humans in ways that seem like a "linear" progressive improvement of our capacities but will ultimately be so different and remarkably capable that its abilities will diverge radically from our own. Superintelligence becomes more incomprehensibly "exotic" the more it gains in power and ability.²¹

The prospect of generating powerful, artificial minds, therefore, entails what Bostrom calls "the control problem" or what Yudkowsky calls "the alignment problem": if superintelligence is so different from us, and is so much "smarter" in ways that increasingly seem alien, how will it be possible to ensure that its "motives" or controlling principles or "values" remain aligned with human interests?

Bostrom spends much of his book *Superintelligence* gaming out scenarios of AI emergence and control. Most of his scenarios entail great risk to human civilisation since there is nothing we can do to prepare ourselves sufficiently for coexistence with

powerful artificial minds. Therefore, Bostrom is rather sober-minded about the trade-offs in cultivating artificial intelligence. Unlike someone such as Ray Kurzweil, who invests “the Singularity” with providential significance, Bostrom tries to reckon honestly and clearly with the possible negative outcomes. He takes this position, in part, because he thinks that our best future lies down the path of AI and artificial general intelligence (“AGI”)²²—that the possible goods are enormous—but that this path needs to be travelled with the utmost care, lest it becomes a civilisational “filter” in the sense borrowed from the Fermi Paradox.²³

Effective Altruism

Effective Altruism (“EA”) judges actions by their outcomes, not intentions. While its focus on effectiveness seems sensible, it masks a commitment to radical philosophical premises. For example, EA posits that proximity or time should not govern other-regarding behaviour.²⁴ In this view, a “util” is a “util” whether it belongs to you, a child in Africa, a person living 17 generations in the future, or even, somewhat controversially, a shrimp. This leads to the conclusion that saving humanity in the infinite future is the most important task for those committed to altruism.

EA’s moral ambition, while inspiring, comes with significant limitations. William MacAskill, a chief theorist of EA, attempts to purify moral goodness, or at least protect “rational morality” from problematic incoherence. However, in his efforts to make morality more effective, he often leaves the impression that genuine moral problems are obviated while overstating the extent to which all obstacles to progress can be reduced to utilitarian trade-offs.

One critical problem for this mode of thinking, and all utilitarianisms, is that it selectively—and unconsciously—rationalises moral intuitions. While EA demands disregarding considerations of time, place, or personal attachment, it uncritically asserts that the *quantity* of suffering matters, as if multiplying pain amplifies its moral “weight”. This raises questions: if reason dictates that proximity should not matter, why should *amount* matter more than *kind*? Moreover, how can we objectively rank or aggregate fundamentally different forms of suffering?

Like rationalism, EA often seeks to escape our natural sociality in favour of “tech-seeming” approaches. This tendency ignores the constraints of our pre-rational elements and attempts to make fungible goods that are sharply heterogeneous for most people, such as the rational search for truth, courageous self-sacrifice in war, or spiritual devotion in religious observance. How do we weigh benefits to ourselves or loved ones against the alleviation of abstract future suffering? What should be preferred in normal times versus extreme cases, like civil wars or the rule of superintelligence? Can reason truly arbitrate or rank these choices? EA’s fundamental premises make these questions impossible to answer.

On the whole, existential risk, with its abstract concern for future life, appears less than rational. It distances us from what we truly are: not disembodied minds but autonomous individuals—body and mind intertwined—whose flourishing is inseparably linked to decentralised communities of friends, family, and communities. It overlooks our nature as beings who navigate complex moral realities, replete with nuanced considerations of loyalty, duty, and personal attachments that resist reduction to commensurable quantification or probabilistic calculations.

When EA meets AI, it produces a problematic equation: even a tiny probability of doom, when multiplied by negative infinity utils (representing the complete loss of all future human potential), equals negative infinity utils. This framework imbues individual actions with cosmic significance, particularly those of technologists. Their ability to manipulate the world through symbols and algorithms paradoxically positions them as potential destroyers and saviours of humanity.

Accelerationism

Unlike the pessimistic school, the proponents of a more techno-optimist approach begin with gratitude for the marvellous achievements of the modern marriage of science, technology, and capitalism.

Accelerationists tend to be liberty-oriented, yet they depart from the moral architecture of the liberal tradition. Marc Andreessen embraces the Hayekian idea of markets as “discovery machines”, revealing and harnessing dispersed, local knowledge through an “exploratory, evolutionary, and adaptive” order.²⁵ This view posits that markets generate abundance by facilitating the increased satisfaction of needs and wants without requiring that humans be angels or the subjects of a superintelligent master planner. Accelerationists share with classical liberalism the view that technological capitalism is the key to human liberation: it allows for diversifying human self-interest into collectively productive ends; it liberates us from violent competition and fear; and it democratises access to leisure.²⁶

However, the extreme version of this optimism regards technology as the key to unlocking man’s potential to overcome himself and become something higher, better, and more refined. The question is not simply one of satisfying our desires but of satisfying them so entirely that we are no longer subject to them.

Technology replaces religion; it is the key to our happiness and salvation. Thus, it should be accelerated by removing impediments to developing and deploying new technology, including AI.

Guillaume Verdon, a leading theorist for accelerationism, argues that life itself is a thermodynamic function seeking to harness free energy. In this view, life finds its natural expression in technological enhancement. Verdon sees no difference, in other

words, between the principles of life and the nature of technology—through the lens of physics, they are one and the same effort to maximise energy access and optimise consumption.²⁷

Verdon and Andreessen give us an accelerationism that departs in important ways from the framework out of which the system of free markets was born. While this school views the “techno-capital” machine as an almost autonomous, self-governing order, classical liberals like Locke and Montesquieu regarded the institutionalisation of natural rights, individual freedom, and limited, constitutional government to be the crucial means for catalysing progress.²⁸ In addition, classical liberalism rests upon a particular conception of what a human being is, namely, a being whose equal right to pursue excellence is confirmed by reason. This foundation produces a vibrant, spontaneous order within civil society based on free, mutually beneficial exchange and association.

The departure from this conception of humanity surfaces a fundamental ambivalence within the accelerationist camp. Some adherents regard human beings as mere products or effects of nature, meaning the thermodynamic laws of the universe. By this view, there is nothing special or dignified about us; therefore, merging with technology to become more fully absorbed as instruments within nature’s sweeping and abstract trajectory is salutary. Passing the baton to AI, as painful or baffling as it may at first seem, is our destiny. This perspective paradoxically views AI as superior to humans, despite being our creation. It is the final embodiment of nature’s directedness towards “an optimisation process” that “[creates] life which constantly expands”.²⁹

The more moderate side of the optimist movement conceives of AI as a potent tool, subordinate to human needs and wants. AI is a “Philosopher’s Stone”, envisioned by Andreessen as having the power to “literally make sand think”, thereby solving a vast array of challenges, from rescuing us from stagnation and boredom to achieving unlimited energy and possibly even overcoming death.³⁰

This dichotomy raises a crucial question: if AI is a mere tool, it will need to be guided by human-defined conceptions of morality and the good, which opens up quite a sphere of deliberation. If it is not a mere tool, but an embodiment of nature’s optimisation process, then we become passive agents of its amoral principles.

The tool-like approach aligns with a so-called “constrained vision”, which is not utopian but incremental. Market signals allow us to shape better tools that satisfy our needs and wants. In contrast, the synthetic, augmented approach envisions a utopian flight from all present discontents and limitations. We become a different kind of being, a “superman” in the diction of the techno-optimist manifesto, echoing Nietzsche; all former questions of morality and politics are irrelevant.

This is why Andreessen presents his philosophy as a “material” and not political philosophy. Its proponents and advocates may be left-wing or right-wing; the parties

of today, shaped as they are by fundamental disagreement over the nature and goals of politics, are irrelevant, as is the tradition of prior thought. This pushing away of the primacy of philosophy to technology goes together with rejecting any dangers or problems associated with technology. And so this shift is one of the temptations of blind optimism: a pollyannaish utopianism that denies, almost *a priori*, that AI could centralise power or sap the creative energies at the heart of Western dynamism.

AI Extremism: A Challenge to Western Dynamism

Notice how, in all of these viewpoints, the human-centric perspective—the world of properly human concerns—is washed away.

Rationalism and long-termism paint a tragic worldview where humans have built the machinery of a civilisation that can no longer sustain itself. They suggest we are either running out of resources or losing the capacity to understand or guide our creations, or both. We are, in essence, the new Prometheans, who, with stolen fire, create idols or gods that become the means of our punishment.³¹ Whatever flourishing we once enjoyed due to our achievements is doomed as we run down the clock on inevitable collapse or bootstrap the technologies that all but guarantee our fleshy obsolescence. AI becomes our daily reminder that we are not gods—gnawing at the liver of human dignity until we expire on the bare rocks of a thwarted future.

From a different angle, it is possible to see a strained hope for cosmic meaning in this pessimism. Like Oppenheimer equating himself to Vishnu in admiration of his genius *and* hubris, the pessimist schools, in stressing the danger of AI, affirm human importance and seek some sort of cosmic response—a kind of Nemesis that will punish us for overstepping our bounds. Something akin to this need is clear in Effective Altruism’s insistence on abstract forms of suffering. Or, to be more specific, it is evident in EA’s compulsion to make us feel that we ought to worry about extremely remote experience and even the suffering of beings who may never be born.

To demand that humans worry in this way, rapt with existential concern about possible futures, far-flung scenarios, and abstracted suffering, resembles the demand that humans show or test their faith by emotionally investing in counter-intuitive notions. The greater the strain on common sense, the better the test. This is the highest and most demanding moral test, like an esoteric dimension of spiritual gatekeeping. It is attractive precisely because it requires a show of commitment, resoluteness, and even a sacrifice of the intellect, all while acting out the appearance of rationality. And it becomes all the more attractive if it makes us feel like we can use “reason” to ward off the Nemesis aroused by our hubris.

Accelerationism is likewise a moral stance, albeit here, the hope is that by focusing on efficiency, we can achieve consistent progress with no adverse consequences. In extreme cases, this hope envisions us becoming god-like—absorbed by The Singularity, rewarded by the cosmos with superhuman abilities, and even granted immortality.

Both pessimism and optimism are driven by a moral hope that urges belief in a reductive and simplified picture of reality. Worse, both threaten to taper off the kind of openness that results from the West's internal dynamism and tradition of self-critique.

While we can sympathise with the desire to find final, rational solutions to technological problems that press us in new ways, especially if these solutions seem conversant with the evident authority of modern science's precision, quantification, and control, we must remain cautious. The Bayesian, the Effective Altruist, and the Thermodynamic Accelerationist each offer us reductive simplicity disguised as highly refined, abstract theory. In all three cases, one conception of reason is elevated above competing views, overshadowing those voices that critique their foundations.

These forms of "solutionism" betray a key caution shared by many of the West's leading minds. From Plato and Aristotle's warnings about the false promises of a political *technē* for the human condition, to Bacon's insistence on purifying science from its idols, to Hume and Nietzsche's concerns about the psychologically motivated distortions of reason, and Mill's worries about utilitarianism, to Kant and Heidegger's critiques of reason itself—the greatest figures in our philosophical tradition consistently urge scepticism toward any simplified vision of reason's pure grasp and control of reality. This scepticism is particularly pronounced when it comes to rational "models" proposed as universal frameworks for policy, ethics, and governance.

What the AI extremists have to offer is thus incompatible with the tension that holds the West together. They seek to resolve this tension—to release the energy in the bowstring—because they have identified our one, true, and final target.

AI extremism wants to close the aperture of the soul, not realising that the tension within the tradition itself—its various competing voices, beliefs, and guiding aspirations—keeps the aperture open and aware of itself.

This lack of self-awareness in AI extremism is especially telling, particularly because the worst experiments in scientistic ideology are so recent in our history. Soviet Marxism tried its best to "solve" the tension of Western dynamism. It, too, purported to have a scientific solution to the West's failures to identify the path of true progress. But like AI extremism, it failed to see how its foreground of "science" and "theory" was motivated in the background by the contortions of suppressed religious hope and moral longing.

The wisdom in Western dynamism is found in its tensions, which make us cautious of singular, definitive solutions, especially ideologies that fail to recognise how their moral aspirations may distort the models they build and attempt to impose, top-down, on entire civilisations.

The failures of current AI theory reveal the need for a different direction and a new approach. This paper will conclude by outlining our positive vision for AI, rooted in a foundational analysis of human flourishing—inspired by the Socratic and Aristotelian understanding of reason and habit, updated with modern research, and demonstrating why reason must work in harmony with certain aspects of the individual mind to support, rather than undermine, our well-being. This perspective aligns with the West’s longstanding focus on individual autonomy as the central principle of its dynamic order and should be the standard for assessing AI’s usefulness.

Our current intellectual climate calls for the cultivation of a new generation of technologists and thinkers whose pedagogy and work are structured by these principles. Following the model set by the newly created HAI Lab at the University of Oxford,³² the goal is to build a future where AI serves human excellence rather than allowing the best aspects of human civilisation to be overshadowed by powerful digital forces, or regulatory policies designed to govern them.³³

The Third Way

We are capable of harnessing the incredible potential of AI by grounding it in an understanding of human behaviour, through the lens of our habits and our tools. By centring these key elements of human experience, we can ensure that AI’s affordances benefit us, rather than allowing the twin dangers of utopianism and over-regulation to eclipse our true interests.

To elucidate this “Third Way”, we will examine four main conceptual elements: habit, tools, offloading, and autonomy. This framework will illuminate where AI can be of enormous use.

Habit

In Aristotle’s *Nicomachean Ethics*, he argues that humans live by their habits, that their “way of being” is rooted in their patterns and regularities. While it might seem banal to say that humans are “creatures of habit”, this very familiar notion conceals far-reaching implications.

In Aristotle’s view, humans not only need habits to live, but require good habits to live well. Habits are structured patterns for organising time, energy, interest, and

behaviour. We often form habits unconsciously, largely because the mental cost of identifying and choosing between alternatives is frequently too burdensome to be worth paying. In essence, humans routinise their lives to avoid the costs of deliberation. This routinisation is necessary because constant deliberation would be paralysing.

The cost of being a self-conscious agent, who is capable of higher order thinking some of the time, is that unthinking patterns take over when it is too costly to be fully aware.

Tools

Aristotle's insights into human nature extend beyond habits to our unique relationship with tools. Unlike other animals, humans are not governed simply by instinct or genetic intelligence. Our behaviour is not wholly determined by our biology. Instead, we possess the remarkable ability to accumulate our experiences as memories, and to employ reason and imagination to picture future outcomes and extract principles from what we have seen, heard, and felt.

These same cognitive capacities that enable habit formation also facilitate our use of tools (*technai*). Humans do not simply live *by* nature, *with* nature, or *alongside* nature; rather, we live by using or actively transforming nature, initially, kinetically, and later, with our artful designs and interventions. For Plato and Aristotle, human life is unthinkable without *technē*. Our technological prowess and labour generate the surplus necessary for economies, which in turn support political communities. The higher pursuits of philosophy and the arts emerge only because of the technological and political substrate.³⁴

Offloading

However, the very aspects of our being that make habits and tools invaluable to us also expose us to a related pitfall. When we offload our thinking to habits or delegate our labour to tools and techniques, we risk becoming overly dependent or trapped in repetitive patterns. While this offloading offers freedom from constant decision-making and the ability to perform important operations without physical effort, it can also become a crutch. In this state of reliance, we fail to perceive alternatives that our habits obscure, and we may neglect to develop the very capacities that our tools have relieved us from exercising. Paradoxically, habits and tools—precisely because they are useful—can blind us to their own side effects and long-term consequences.

For Plato and Aristotle, the costs of offloading require us to remain vigilant about the ultimate goals of life, which both philosophers identify as “virtue” or “the good”. It is a constant battle, as it were, to cultivate the framework in which habits and tools can support a good or decent life. Because we know great unhappiness can stem from

harmful habits or states of dependence, our reason tells us that we should cultivate the rational faculties for governing our habits and intervening when in fact they undermine our flourishing.

Therefore, habits and tools are susceptible to corruption; they do not naturally or spontaneously organise themselves for our benefit. This makes sense, as habits and tools fill the gap created by our separation from nature. Unlike other species, humans are not automatically governed by instinct or strictly guided by the script of DNA.

Plato and Aristotle rightly thought that regimes rise, decline, and collapse in a cycle driven by early success, which often gives way to later decadence.³⁵ The historical failure of civilisations to solve the problem of endurance serves as further evidence that habits and tools, being disconnected from rational reflections on our flourishing, do not inherently lead to linear progress.

Autonomy

Because humans enjoy a remarkable degree of freedom from nature, supported by technology, politics, and the economy, it is tempting to believe that technology alone can fully govern the ends of human life—the goals that shape our habits and tools.

This temptation is not new. In the ancient world, it dawned on the Greeks with the prospect of science and scientifically enhanced *techne*. Several thinkers—philosophers and sophists—thought it possible to develop an art of politics, believing that the same principles we apply to farming or hunting might be used to organise human beings into stable configurations.

Plato and Aristotle rejected this idea, arguing that creating a perfectly just or stable society—one that eliminates all risk or predictably satisfies every citizen's desires—is impossible without undermining individual striving for excellence. This is not to suggest that there is nothing *techne*-like to learn from politics, but rather that any principles we derive will fall short of fulfilling our hopes for a totalising order.³⁶

This insight from Socratic philosophy harmonises with both the Homeric and Biblical visions. All three affirm, in their own ways, the need to protect the individual from the temptation to close our political horizons by elevating some idol, ideology, regulation, or *techne* to ultimate authority.

Instead, the West has embraced the idea that humans should be free to think about their own lives, protected from violent or prescriptive compulsion. This approach suggests that the best way to align political governance with our reliance on habits and tools is to give people the freedom to experiment with their lives—to consider how they want to live, how to cultivate their own talents, and how to govern their use of *techne* in ways that support their autonomy in pursuit of excellence.

Thus, autonomy finds its foundation in what we call liberal education and liberal governance—fostering inner freedom and the cultivation of mind and insight, alongside protections from coercion and external influence.

AI and Human Autonomy

The rise of AI presents new challenges to human autonomy. Digital technologies like smartphones and social media, now being supercharged by AI, seamlessly integrate with our habits and reshape our routines. The impact is already significant: an estimated 20% of our discretionary time is currently mediated by algorithmic information feeds. This influence will only intensify as AI becomes embedded in all mediating technologies over time.

AI's unique power lies in its versatility. In many ways, AI is the *tool of tools*, the generalised problem-solver, the multi-tasking, multi-dimensional pocket-knife for the mind.

This power, however, is a double-edged sword. AI offers us total solutions, positioning itself as a solver of life's problems. We talk to it and let it talk to us. We interface with it at the level of habit and choice. We use it like a GPS for life, an all-purpose choice-maker or life-algorithm.

Because AI integrates so seamlessly into our lives, it can exploit the blind spots we develop through our reliance on habits and tools. Yet, paradoxically, this is where AI can be especially valuable. Imagine tools that we could train to help us avoid our most common mistakes, so we could make new, more informative ones, or that expanded our perspectives rather than narrowing them. Consider tools that could bring to life multiple viewpoints and actively engage with the different sides of the debates and decisions that shape our lives. That could begin from where we are and push us beyond ourselves in ways consistent with our flourishing. In short, tools that could help us reason *better*.

AI already greatly improves the efficiency of our tools and systems. But for AI to remain helpful in this domain, it must be governed by the openness that protects our autonomy and preserves the soul's aperture on the world.

To achieve this means using AI to cultivate better inquiry at scale.³⁷ For example, HAI Lab at the University of Oxford is using concepts of reason from Plato's *Theaetetus* to guide experiments on AI "thinking", aiming to better understand how AI can be informed by and aligned with the question-driven structure of human thought.

In addition, we should advocate for AI that amplifies dispersed, local intelligence. Communities can train AIs to explore solutions that have emerged from their cultures and traditions—the ways their respective cultures have traded off the costs and

benefits of habit, tool use, market forces, moral norms, and governance. We believe that competing, AI-amplified governance models are the way forward, not top-down world states that impose one blanket order for all.

To resist the temptation to use AI to “solve” human life or government or civilisation, we ought to have different *houses of sensemaking*, perhaps working alongside academia and institutional science, that organise new and old research around competing philosophic perspectives.

Ultimately, in all of these efforts, AI could become the tool that makes tools more efficient, while also protecting us from the blindness we risk as creatures of habit and tool-use. In this way, AI could be one of the greatest allies of human autonomy, with applications that augment the main supports of our freedom: tools, markets, science, and governance.

As a civilisation, we must aim to cultivate a network of thinkers and builders who, in their own unique ways, experiment with the intellectual and cultural resources of the West to usher in an AI future that promotes human flourishing. As we have seen, philosophy plays a crucial role in this effort—not only in critiquing competing schools of thought on AI, but also in helping us engage with the fundamental questions we must address to safeguard our autonomy from our dependence on technology, as well as from the constraints imposed by regulation, ideology, and other systems of top-down control.

To harness AI's potential for human good, we must ensure that we are not overwhelmed either by the machines themselves or by the narratives we construct to control them.

¹ Os Guinness, "Our Civilisational Moment", *ARC Research*, 2023, p. 9, <https://www.arc-research.org/research-papers/our-civilisational-moment>.

² Jack Clark, "Import AI 337: Why I am confused about AI; penguin dataset; and defending networks via RL with CYBERFORCE", *Import AI*, (21 August 2023) <https://importai.substack.com/p/import-ai-337-why-i-am-confused-about>.

³ The "alignment problem" refers to the difficulty humans may encounter in aligning the behaviour of AI agents with human morality and values. The alignment problem is underpinned by Nick Bostrom's "orthogonality thesis", which he defines in his book, *Superintelligence*. It states, "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." If AI tools are intelligent enough to act as agents in performing instrumental actions towards final ends, then it is possible that AI agents will be willing to perform actions or pursue ends that are incompatible with human morality. Relatedly, the term "p(doom)" is shorthand for "probability of doom", and refers to the chances that AI produces a "doomsday event", perhaps because humanity fails to adequately address the alignment problem. Various leaders in AI technology publicise their opinion of "p(doom)" as a percentage, from 0 to 100. See Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2016. See also Sean Thomas and Brendan O'Neill, "Are we ready for P(doom)?", *The Spectator*, 4 March 2024, <https://www.spectator.co.uk/article/are-we-ready-for-pdoom/>.

⁴ Nick Bostrom, "The Vulnerable World Hypothesis", *Global Policy*, vol. 10, no. 4, 2019, pp. 455–476, <https://nickbostrom.com/papers/vulnerable.pdf>; Future of Life Institute, "Pause Giant AI Experiments: An Open Letter", *Future of Life Institute*, 22 March 2023, <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

⁵ Ray Kurzweil, *The Singularity Is Near*, Viking, 2005; Marc Andreessen, "The Techno-Optimist Manifesto", *Andreessen Horowitz*, 16 October 2023, <https://a16z.com/the-techno-optimist-manifesto/>.

⁶ Aristotle, *Nicomachean Ethics*. See Book III for a treatment of rational deliberation's relation to choice and voluntary action (1112a–1113a), Book VI for a treatment of rational deliberation's relation to prudence (1140a–b), and Book VII for rational deliberation's relation to self-mastery, through self-restraint.

⁷ Homer, *The Iliad*; Exodus; Daniel 6:16–23; Plato, *The Republic*; Aristotle, *Politics*; John Locke, *The Second Treatise of Government*.

⁸ This intellectual crucible of Western dynamism has forged many of America's greatest leaders and thinkers: James Madison's immersion in the history of statecraft, Abraham Lincoln's engagement with Shakespeare and the Bible, J.P. Morgan's religious influence on Teddy Roosevelt, Frederick Douglass's self-education and championing of self-reliance, Walt Disney's futurism, Martin Luther King Jr.'s theological appeals, and Dorothy Day's compassionate activism. If a nation's greatness is measured by its individuals, then America's faith in its citizens to define their own divergent paths has been one of history's most successful experiments.

⁹ Brendan McCord and the Cosmos Institute, "Existential Pessimism vs. Accelerationism: Why Tech Needs a Rational, Humanistic 'Third Way'", *Cosmos Institute*, 4 January 2024, <https://cosmosinstitute.substack.com/p/existential-pessimism-vs-accelerationism>.

¹⁰ Bayesianism is an ideology which seeks to mathematise rationality by assigning to various events or beliefs probabilities that can be manipulated using statistical methods to arrive at seemingly rational, precise, and unbiased conclusions. In general, Bayesianism equates the predictive power of probabilistic techniques with rationality, itself, and sometimes with causal explanation. The Bayesian statistical method *par excellence* is Bayes Theorem, which is a method for updating the probability of an event, given a prior probability of that event and freshly observed evidence. Bayesianism's cold and calculating self-image is in significant tension with the fact that "priors" for complex situations are often derived from intuitive judgement calls. Furthermore, detractors of Bayesianism will point out that real life is often too messy to use law-like methods appropriately, and that overuse of probability distributions often hides insufficient justification for espousing beliefs, resulting in ironically biased or arbitrary conclusions. See Eliezer Yudkowsky, "Beautiful Probability", *LessWrong*, 14 January 2008, <https://www.lesswrong.com/posts/bkSkRwo9SRYxJMISY/beautiful-probability>. See also Ben Landau-Taylor, "Probability Is Not A Substitute For Reasoning", *Ben Landau-Taylor*, 21 March 2023, <https://www.benlandautaylor.com/p/probability-is-not-a-substitute-for>.

¹¹ Professor Philipp Koralus, Director of the Human-Centered AI Lab at the University of Oxford, extends this longstanding tradition by proposing a theory of reason that emerges not from probabilistic computation but from the progressive clarification of the right questions.

¹² The Less Wrong community, founded by Eliezer Yudkowsky in 2009, is an online forum dedicated to discussing rationality, cognitive biases, and artificial intelligence. It popularised concepts like "existential risk" and "AI alignment", influencing debates on the potential impacts of advanced AI. The community's ideas have gained traction in Silicon Valley and among some AI researchers, shaping discussions on the future of technology and human decision-making. <https://www.lesswrong.com/>.

¹³ Eliezer Yudkowsky, "The Only Way to Deal With the Threat From AI? Shut It Down", *Time*, 29 March 2023, <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough>.

¹⁴ Future of Life Institute, "Pause Giant AI Experiments: An Open Letter", *Future of Life Institute*, 22 March 2023, <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

¹⁵ Eliezer Yudkowsky, "The Only Way to Deal With the Threat From AI? Shut It Down", *Time*, 29 March 2023, <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough>.

¹⁶ *Ibid.*

¹⁷ Ari Schulman, "Open Wallets, Empty Hearts", *The New Atlantis*, Summer 2024, <https://www.thenewatlantis.com/publications/open-wallets-empty-hearts>.

¹⁸ Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards", *Journal of Evolution and Technology*, vol. 9, March 2002, <https://nickbostrom.com/existential/risks.pdf>.

¹⁹ Plato, *Republic*, Book VIII. Plato, *Laws*, Book III.

²⁰ Nick Bostrom, "The Vulnerable World Hypothesis", *Global Policy*, vol. 10, no. 4, 2019, pp. 455–476, <https://nickbostrom.com/papers/vulnerable.pdf>.

²¹ Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2016.

²² Artificial general intelligence is distinguished from artificial intelligence by a wider breadth and higher degree of intelligence. AI is widely accepted to exist already, in technologies like general pretrained transformers ("GPT"). On the other hand, it is widely accepted that AGI does not exist and would involve an intellect that equals human cognitive performance across all domains. Nick Bostrom speculates, "If general AI were to be achieved, this might also lead to superintelligence", i.e., "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest". See Vincent C. Müller and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion", *Fundamental Issues of Artificial Intelligence*, 2016, <https://nickbostrom.com/papers/survey.pdf>.

²³ Nick Bostrom, "Where Are They? Why I Hope the Search for Extraterrestrial Life Find Nothing", *MIT Technology Review*, May/June issue (2008), pp. 72–77, <https://nickbostrom.com/papers/where-are-they>.

²⁴ Peter Singer, "Famine, Affluence and Morality", *Philosophy and Public Affairs*, Vol. 1, No. 3 (Spring, 1972), pp. 229–243.

²⁵ Marc Andreessen, "The Techno-Optimist Manifesto", *Andreessen Horowitz*, 16 October 2023, <https://a16z.com/the-techno-optimist-manifesto/>.

²⁶ Ibid.

²⁷ Guillaume Verdon, "Notes on e/acc principles and tenets", *Beff's Newsletter*, 9 July 2022, <https://beff.substack.com/p/notes-on-eacc-principles-and-tenets>. Note that Verdon published various writings under the pseudonym "Beff Jezos" and co-authored this piece with another pseudonymous author, "Bayeslord".

²⁸ John Locke, *Second Treatise of Government*, 1689. See Chapter II, Of the State of Nature, and Chapter V, Of Property.

²⁹ "Effective Accelerationism", <https://effectiveacceleration.tech/>. Note that, at the time of this writing, the webpage referenced is defunct. An archive is available here: <https://web.archive.org/web/20231121002029/https://effectiveacceleration.tech/>.

³⁰ Marc Andreessen, "The Techno-Optimist Manifesto", *Andreessen Horowitz*, 16 October 2023, <https://a16z.com/the-techno-optimist-manifesto/>.

³¹ Toby Ord, *The Precipice: Existential Risk and the Future of Humanity*, Hachette Books, 2020. See discussion of "unrecoverable dystopias" in Chapter 5: Future Risks.

³² See the website for the Institute for Ethics in AI, accessed 3 January 2025, <https://hailab.ox.ac.uk/>.

³³ "Oxford launches Human-Centered AI Lab", University of Oxford, 5 September 2024, <https://www.ox.ac.uk/news/2024-09-05-oxford-launches-human-centered-ai-lab>.

³⁴ Plato, *Republic*. Aristotle, *Politics*.

³⁵ See Plato's *Republic*, Book VIII, for a description of cycles of political regimes, involving the sequential decay of aristocracy into timocracy, oligarchy, democracy, and finally tyranny (544a–569c). See also Aristotle's *Politics*, Book V, for a treatment of the rise of factions in various regime types, and the types of regime that are likely to result from the subsequent decay (1301a–1307b). Note, however, that decadence does not necessarily *cause* societal collapse. Following Tainter, a central cause of societal collapse may be declining returns on increased civilisational complexity, e.g., technologies, institutions, and other specialised systems. See Joseph Tainter, *The Collapse of Complex Societies*, Cambridge University Press, 1988.

³⁶ Plato, *Republic*. Aristotle, *Politics*. Leo Strauss, *The City and Man*, University of Chicago Press, 1978.

³⁷ Philipp Koralus, *Reason and Inquiry: The Erotetic Theory of Reasoning*, Oxford University Press, 2020.



This research has been produced and published by ARC Research, a not-for-profit company limited by guarantee registered at Companies House with number 14739317, which exists to advance education and promote research.

ISBN: 978-1-916948-47-1

www.arcforum.com

February 2025