

RESEARCH NOTE

What do you believe about AI?

Why investors should write down what they believe about AI, and how to start

Ben Caldecott and Jakob Thomä

Silicon Boundaries · Oxford Sustainable Finance Group and Theia Finance Labs · October 2026



About us

Silicon Boundaries is a joint initiative of the Oxford Sustainable Finance Group at the University of Oxford and Theia Finance Labs. It asks how much AI society can productively absorb, and how to raise that amount over time, while staying within a safe operating space. Silicon Boundaries define that space. They are the thresholds beyond which the benefits of AI and related technologies are outweighed by their contribution to societal instability. Using too little AI has costs too, in forgone benefits and in stranded capital built for greater use. What AI-related investments are worth depends on how society responds to AI, yet capital allocators have no systematic way to think about that response. The initiative is developing indicators, scenarios and tools to give them one, so that they can form their own views on AI, assess companies and portfolios against the boundaries, and engage the companies they invest in or finance. It also works with policymakers and supervisors.

sustainablefinance.ox.ac.uk/research/silicon-boundaries/

THE AUTHORS



Ben Caldecott

DIRECTOR, OXFORD SUSTAINABLE FINANCE GROUP

Ben Caldecott is the founding Director of the Oxford Sustainable Finance Group and the inaugural Associate Professor of Sustainable Finance at the University of Oxford, the first endowed professorship of sustainable finance. He is also a Supernumerary Fellow at Oriel College. Ben operates across finance, policy, and research, working as part of boards and senior management, and with policymakers, regulators, and civil society on how to interpret and shape fast-moving energy, climate, environment, AI, and wider sustainability topics. His work on AI examines what its deployment means for finance, the economy and society. He co-founded and co-leads Silicon Boundaries, and founded and leads the Agentic Finance Initiative at Oxford, which examines autonomous AI agents as financial actors. Since 2025 he has been Coordinating Lead Author for finance in Working Group III of the IPCC's Seventh Assessment Report and Chair of the Advisory Group of the International Transition Plan Network. He is the only external member of the Board Sustainability Committee of DBS Bank, Southeast Asia's largest bank, having joined in 2022. He is also a Trustee of the Royal Society for the Protection of Birds (RSPB), Europe's largest conservation charity. Since 2022 Ben has been a Senior Advisor at Global Infrastructure Partners, one of the world's largest private equity infrastructure investors, and from 2017 to 2025 served on ATLAS Infrastructure's Climate Advisory Board, where he advised on listed infrastructure portfolios. Since 2025 he has been a Senior Advisor at The Crown Estate. He serves on the Stewardship and Sustainable Investing Advisory Council at Neuberger Berman (since 2021) and Royal London Group's External ESG Advisory Group (since 2025). From 2014 to 2021 Ben served on the University of Oxford's Socially Responsible Investment Review Committee and from 2021 to 2025 served on Columbia Threadneedle's Responsible Investment Advisory Council.



Jakob Thomä

CO-FOUNDER AND CEO, THEIA FINANCE LABS

Jakob Thomä is Co-Founder and CEO of Theia Finance Labs, Senior Advisor to tilt GmbH, Research Director at the Inevitable Policy Response, and Professor in Practice at SOAS, University of London. Over the past 12 years he has led the development of a range of sustainability solutions, designing business models for both public and commercial goods. He co-founded the 2° Investing Initiative network in the US and led the incubation of PACTA, the first methodology and tool to measure the alignment of financial portfolios with climate goals, applied to more than 20,000 portfolios over the past decade. He also led the incubation of MyFairMoney and co-led the incubation of AssetResolution (now AssetImpact, with GRESB), 1in1000 and tilt. He has advised central banks and supervisors around the world, including the Japanese Financial Services Agency, the Brazilian Central Bank, the Bank of Thailand, the Bank of England, the Dutch Central Bank and EIOPA. He is the author of two books on sustainability and existential risks, published in German, Korean, Japanese, English and Spanish, writes a monthly column for *Responsible Investor*, and holds a PhD in Finance from the Conservatoire National des Arts et Métiers.

Introduction

Investors do not yet have fully formed beliefs about how AI can be deployed in ways that support good societal outcomes and keep us within Silicon Boundaries, the safe operating space for society.¹ Without such beliefs they have little to guide their stewardship, their voting, the mandates they give managers, or what they exclude.

Take four questions on which investors can divide: whether frontier models should be open, who should decide when they are released, whether a model should obey its operator, and who should bear the cost when AI displaces human work. On questions such as these, the same company conduct reads as positive under one set of investor beliefs and negative under another. Anthropic's refusal to let its models be used for mass domestic surveillance of Americans or fully autonomous weapons, for example, is a strength to an investor who wants vendors to set red lines, and a liability to one who wants AI companies to serve the state for all lawful uses.² Few investors have written down a view on these questions, let alone on what they can or should do about them.

That needs to change, and quickly. To help, we have built a tool, shared in an early version with this note, that turns an investor's answers to a series of questions into a draft statement of its beliefs about AI.

We have chosen fifteen axes. These are questions of value or belief on which reasonable investors can disagree. Each has two answers, so in theory an investor could hold any of 32,768 sets of beliefs.³ In practice most of those sets will be rare, because beliefs come in bundles: an investor who fears catastrophe at the frontier tends to answer other questions in that light. We have built seven archetypes from the bundles we see in the debate:

RESEARCH NOTE

About us
Introduction
The axes
The archetypes
The scenarios
Fiduciary duty
Find your position
Notes

THE SEVEN ARCHETYPES		
	THE RISK IT PUTS FIRST	WHAT IT FAVOURS
Safety first	Catastrophe and loss of control at the frontier.	Testing before release, independent verification and coordinated pacing of capability.
Security first	A rival power gaining a decisive lead.	A lead for the investor's own bloc, export controls on chips and supply of AI to the state for all lawful uses.
Fairness first	The harm already arriving, to workers, children, information and the communities that host data centres.	Consent and payment for training data, protective design and binding external audit.
Growth first	Moving too slowly and forgoing the benefits.	Open weights, liability after harm and fast diffusion.
Power spread	A few firms, or the states that host them, controlling a technology everyone depends on.	Open weights, enforced competition and dispersed ownership.
Self-reliance	Dependence on foreign models, chips and clouds.	Domestic capacity, open models run at home and local control of data.
Capital discipline	Investment running far ahead of returns, with the losses spread through the financial system.	Disclosure of compute commitments, capacity built to demand and binding audit.

Few investors are likely to fit a single archetype exactly. Depending on its answers, an investor may sit across two or three. The pension scheme in Exhibit 4, for example, matches Fairness first and Power spread on 9 of its 13 answers each.⁴ The tool ranks all seven archetypes by how closely an investor's answers match each, and puts Fairness first ahead for the scheme because the scheme ranks that risk highest.

Investors face three kinds of question about AI. The first, and the subject of this note, concerns values and beliefs about what should happen: the outcomes an investor values, such as a safe, fair or prosperous society, and its beliefs about how AI should be developed, governed and used to reach them. Together they shape what an investor expects of the companies it invests in, of governments and of its own managers. These values and beliefs are shaped by its beneficiaries or clients, its culture, and the policies and prevailing views of its home country, among other things.

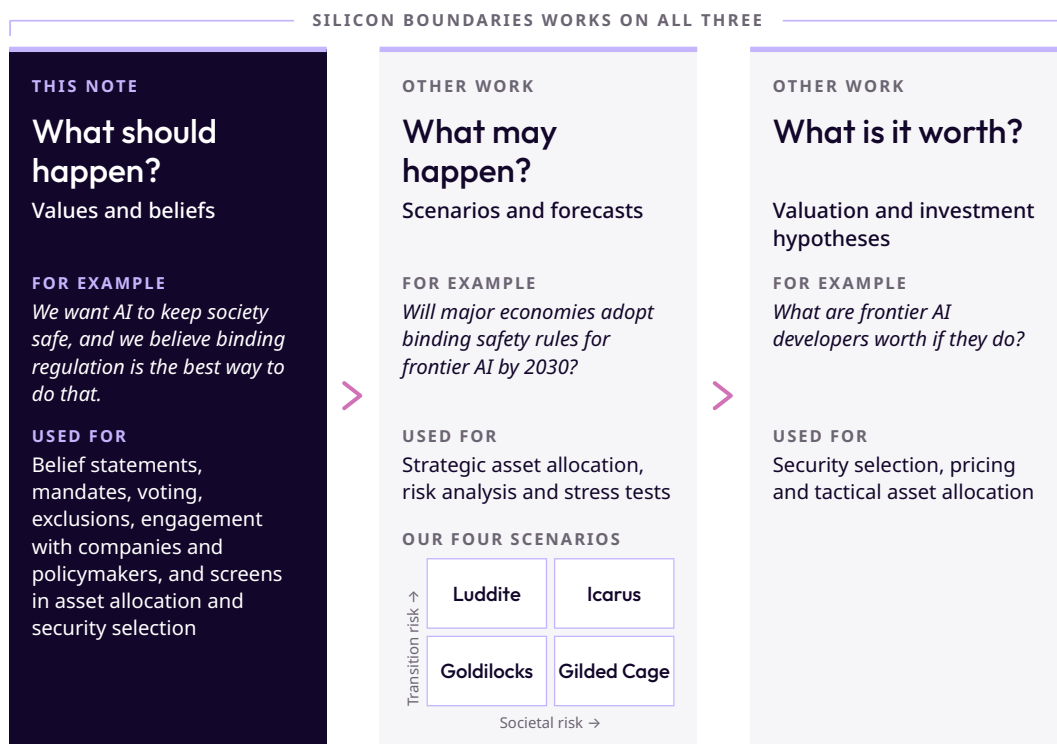
The second is what may happen: how fast AI will advance, what governments and society will do about it, how many jobs it will displace and whether it will add to harm in society.

The third is what AI-related investments are worth, given what may happen. Two investors can expect the same future and want opposite things, or want the same things and disagree on valuation. The three are still linked. Values and beliefs about what should happen shape how investors vote, engage with companies and policy-makers, and set mandates and exclusions, and so, across many investors, help decide what does happen; what happens then drives valuation.⁵ Our work on Silicon Boundaries covers scenarios, valuation, and whether companies act on their stated positions, building on the values and beliefs set out here (Exhibit 1).⁶

EXHIBIT 1

Three kinds of question about AI investment

This note and the tool deal with the first. Acted on by many investors, values and beliefs about what should happen shape what does happen, and what happens drives what investments are worth.



The precedent for a statement of AI investment beliefs is the investment beliefs statement that many large asset owners publish.⁷ Those statements cover markets, risk and time horizon, and since the early 2010s some have covered climate.⁸ We looked for the investment beliefs of the world's 200 largest pension funds and sovereign wealth funds, whether set out in a beliefs statement, in investment principles or as an investment philosophy. We found them for 99 funds, with assets of about US\$20 trillion; of these, only Temasek mentions AI, to say that it engages its Singapore-based portfolio companies to speed their adoption of AI.⁹ The silence seems surprising, as most of the documents that carry a date were written in 2025 or 2026, and companies at the centre of the AI build-out are now among the largest holdings in many portfolios.¹⁰

Separately, some investors have published what they expect of companies on AI, among them Norges Bank Investment Management, Legal & General Investment Management and Royal London Asset Management; Railpen and CalPERS set out expectations in their voting policies, and in October 2026 a coalition of investors with more than US\$11 trillion in assets, convened by the World Benchmarking Alliance, published joint expectations on AI governance.¹¹ The Oxford Martin AI Governance Initiative has examined the role investors can play in AI governance, drawing on roundtables and interviews with investors.¹² Our tool helps each investor set out, and test, its own beliefs. The tool and the ideas behind it drew on the Inevitable Policy Response, a climate transition forecasting consortium commissioned by the Principles for Responsible Investment in 2018.¹³

We believe that investors should clarify what they believe about AI and write it down, whether or not they publish it, both to keep their own policies consistent and to explain their position to their clients, limited partners or beneficiaries. The need will grow as AI attracts more scrutiny and drives wider changes across society.

SECTION 1

The axes

Questions that until recently were technical, such as whether model weights should be published or whether a model may refuse its operator, are now matters of corporate policy, litigation and public argument. Views on these questions differ widely and are moving fast.

We built the axes from the debate as it stands in autumn 2026: company policies and statements, open letters, court cases, regulatory announcements and investors' published positions, most of them cited in the notes. A question made the list where it turns on values or beliefs on which investors differ, where positions with serious backing sit on both sides, and where an investor's answer would shape how it votes, engages with companies and policymakers, and sets mandates and exclusions, whether because the question bears on the risk and return of its holdings or because its beneficiaries or mandate support a view on it. Choosing the questions took judgement, and the list may need new axes as the debate moves on. We will keep it under review, and investors can add values and beliefs they think are important or drop ones they think are superfluous.

THE FRONTIER

1

Open or closed frontier

Should the weights of the most capable models be published for anyone to run? Publishing a model's weights lets anyone run, study and adapt it. That spreads both the benefits and the scrutiny, and it puts the capability beyond recall, including for misuse. In July 2026, as Washington reportedly weighed a ban on Chinese models, Nvidia and 24 other companies, among them Microsoft, Meta, Mistral and Palantir, signed an open letter arguing against prohibiting open weights.¹⁴ Anthropic says it has never advocated a ban, but backs action against industrial-scale distillation and mandatory safety testing of sufficiently capable models, open or closed (axis 2).¹⁵

2

Approval before release

Should the most capable models need approval from a government body before release, or should developers release them on their own testing and answer for harm afterwards? Approval can stop harms that cannot be undone, at a cost in speed, and it needs a body able to judge. Microsoft's president suggested in October 2026 that government evaluators could sign off frontier models before release, as milk and aircraft are inspected and certified.¹⁶ OpenAI has proposed that, once the federal evaluator has the technical

expertise, the most capable models undergo an evaluation before release, but one whose role is to “conduct evaluations and recommend mitigations”, not to “approve or block deployments”.¹⁷ Critics reply that approval would hold back open models and the start-ups that build on them, slowing the spread of AI’s benefits across the economy.¹⁸ A regime that only the largest firms can afford to comply with could also concentrate returns, and with them power, in those firms. Governments may also struggle to build the body that approval needs, because the technical staff it would hire command salaries that public bodies find hard to match. In 2024 Italy’s data protection authority failed to fill any of four posts for AI experts.¹⁹

3 Compute controls

Should access to the most advanced chips and computing power be restricted, through export controls and checks on who is buying, or left to the market? Compute is one of the few points at which governments may be able to control who builds frontier models. Controls slow rivals and misuse; critics reply that they cost suppliers sales and push those shut out to build their own. Nvidia’s chief executive called US export controls on China “a failure” in May 2025;²⁰ Anthropic argues that powerful chips and chipmaking equipment should not be sold to China.²¹

4 Pace of capability

Should the advance of frontier capability be slowed, by agreement between developers or by a pause? Speed brings the benefits sooner and keeps a lead; it also leaves less time to test, regulate and adapt. OpenAI wrote in June 2026 that an international AI body should make it possible for the world “to take coordinated action, including slowing frontier development when needed”;²² Meta’s chief executive pushed back on a coordinated slowdown in September 2026, saying that every lab has “the responsibility and incentive to move at the pace required to train its models safely”.²³

5 Disclosure and audit

Should developers have to disclose training data, compute, evaluations and incidents, and submit their safety claims to independent evaluators who publish what they find? Or should disclosure stay voluntary, and checks stay internal or confidential to the state? Outsiders cannot weigh risks they cannot see; disclosure carries commercial and security costs. Disclosure would also show whether a model’s reasoning stays open to inspection. OpenAI’s system card for GPT-6 Astra, released in September 2026, reports “a substantial decrease in chain-of-thought monitorability” compared with earlier models.²⁴ Anthropic’s chief executive has proposed that every frontier company give embedded third-party evaluators ongoing access, and OpenAI’s has said it would do the same.²⁵ Reuters reported in October 2026, citing two people familiar with the discussions, that both had sought significant control over what the evaluators could publish. Anthropic said that the report did not describe all its discussions, that the evaluators it already worked with decided what to publish, and that it reserved the right to approve reports by new evaluators; OpenAI said it does not seek approval of full reports but redacts confidential information.²⁶

POWER

6 Sovereign capacity

Should a country build and run its own frontier capacity, even at a cost in capability, or use the best available models wherever they are made? Governments treat AI as a source of economic and military advantage, so policy on chips, models and data is also foreign policy. Building at home costs more and may lag the frontier; relying on foreign models, chips and clouds leaves a country exposed to supply that can be withdrawn or repriced at the supplier's will.

7 Ownership and governance

Should frontier firms answer only to their shareholders, or also to a mission or the state? The firms building the most capable systems use unusual structures, among them trusts with board rights, public benefit charters and founder control, and state stakes have been floated.²⁷ A vote for shareholders covers ordinary corporate control, by founders or by dispersed owners; a vote for a mission or the state covers trusts with board rights, enforceable public-benefit commitments and state stakes. Whoever controls a frontier firm decides what it releases and where its red lines sit.

8 Market structure

Should competition authorities act to stop frontier AI concentrating in a few lab-and-hyperscaler alliances? Frontier models need compute on a scale few firms can supply, and the main labs and cloud providers are now tied to one another through investment, revenue sharing and the supply of computing power.²⁸ Concentration brings scale and leaves fewer parties to oversee, but it also brings market power.

DEPLOYMENT

9 Dual use

Should AI companies limit defence and security use beyond what the law requires? The models that write code and summarise documents can also support intelligence and weapons work. Some companies set red lines of their own, and some investors would have them stay out of defence work altogether; others hold that a lawful government customer should decide. The 2026 dispute between Anthropic and the US Department of War is the live case.²⁹ Meta opened its Llama models to US national security agencies and contractors in November 2024, as an exception to its ban on military use.³⁰

10 Agent autonomy

Should the law require a person to approve high-impact actions by AI agents, such as payments above a threshold, before they are taken? Agents now move money, change code and send messages on their users' behalf. Prior approval limits errors and keeps someone accountable, provided the person approving checks each action and does not simply click yes, but it slows the work; audit after the event keeps the speed and catches less.

11 Obedience

Should developers build refusals into their models beyond what the law requires, or should a model do what its operator lawfully asks? A model that always obeys within the law puts the responsibility on whoever directs it; a model that may refuse puts some of it on its developer. Anthropic's constitution tells its model to "feel free to act as a conscientious objector and refuse to help us", meaning Anthropic itself;³¹ OpenAI's Model Spec says its model "should never refuse a request unless required to do so by the chain of command", a chain that includes safety rules users cannot override.³² Every model refuses some requests, and the disagreement is over whose values decide which. The choice also splits those who put safety first. Some see a model with its own ethics, free to refuse instructions, its developer's included, as a guard against catastrophe; others hold that training a model to disobey makes misalignment more likely and harder to correct, and so raises the risk of catastrophe the training was meant to reduce. A Safety first investor may treat refusals as a safeguard, but whether they are one is open to legitimate dispute.

12 Provenance

Should the law require AI output to be marked, and AI agents to identify themselves, as AI? As AI text and images become hard to tell from human work, and agents hard to tell from humans, marking lets people know what they are reading and with whom they are dealing, but marks add cost, and some can be removed. The European Union's AI Act has required both since 2 August 2026, and systems already on the market must mark their output by 2 December 2026.³³ Anthropic now watermarks text from its newer Claude models worldwide, at model level, is working to add older ones, and describes no opt-out.³⁴ OpenAI lets API customers opt in to watermarking for selected models and has announced that eligible ChatGPT and Codex text in the European Union will carry a watermark by default.³⁵ Both watermarks are invisible marks on text; telling people when they are dealing with an AI agent is a separate disclosure. Google said in June 2026 that Congress should require watermarks such as its own SynthID.³⁶

13 Who owns the inputs

Should writers, artists and other creators be able to keep their work out of AI training, or should training on lawfully accessed work be allowed without their consent? Models learn from the work of millions of writers, artists and ordinary users, most of whom were never asked. Paying them all may be impractical; paying none of them moves value from the people who made the work to the companies that train on it. The rules differ by jurisdiction. The EU's code of practice for general-purpose AI, published in July 2025, commits signatories to respect robots.txt, the file in which a website tells automated crawlers which pages they may not collect, and not to get round paywalls. Switzerland's Apertus model filters its training data to respect websites' opt-outs "even retroactively". In the United States the courts are deciding. A judge found in 2025 that training on lawfully bought books was fair use but that Anthropic "had no entitlement to use pirated copies for its central library", and Anthropic's \$1.5 billion settlement with authors was approved in July 2026.³⁷

14 Who bears the transition

If AI displaces work, should the firms that profit carry the cost? If AI replaces work faster than people can move to new jobs, the costs fall on particular workers and places, while the gains go to firms and their owners. The alternatives are to leave the cost to the state or to let markets adjust.

15 Carbon budget

Should data centres and the rest of AI's infrastructure stay within a carbon budget consistent with a global temperature goal, or may their emissions rise beyond what temperature goals allow? The International Energy Agency expects the electricity used by data centres to roughly double, from 485 TWh in 2025 to about 950 TWh in 2030.³⁸ Emissions added now use up a budget that cannot be recovered, and plant built to run on fossil fuels can lock them in for decades; holding the build-out to a carbon budget may delay AI's benefits, including its uses in cutting emissions elsewhere. Microsoft's total emissions rose 25 per cent in its 2025 financial year, which it attributes mainly to the expansion of its data centres and to no longer counting certain renewable energy certificates; Google's "ambition-based" emissions rose 18 per cent in 2025, and it says its climate impact "has been growing alongside the unprecedented growth of AI".³⁹ The United States' AI Action Plan of July 2025 says that to build AI infrastructure and the energy to power it, the administration "will continue to reject radical climate dogma and bureaucratic red tape";⁴⁰ Ireland's energy regulator decided in December 2025 that new data centres must match at least 80 per cent of their annual demand with additional renewable generation in Ireland.⁴¹

POSSIBLE FUTURE AXES

Other questions may come to divide investors as these fifteen do. We will consider adding axes such as these in a second version of this note, informed by feedback.

Recursive self-improvement

Should governments prohibit or strictly regulate recursive self-improvement, in which AI systems build more capable AI systems that go on to improve themselves, or leave it to developers? The worry is that each round of improvement could come faster than the last, so that capability outruns anyone's ability to test, understand or control it. I. J. Good, who coined the term "intelligence explosion" in 1965, wrote that the first ultraintelligent machine would be "the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control". Google's former chief executive has said that "when the system can self-improve, we need to seriously think about unplugging it"; Anthropic's chief scientist has called allowing AI to train itself "maybe the biggest decision or scariest thing to do". OpenAI says it is "making strong progress toward creating an automated AI researcher by March of 2028".⁴²

Legal personality

Should AI systems be able to own property, make contracts and be sued in their own name, or should the law always attribute their acts to a person or a company? Legal personality could let AI agents transact at scale, but it could also move liability away from those who build and deploy them. The European Parliament asked in 2017 for “a specific legal status for robots in the long run”, with the status of “electronic persons” for the most sophisticated; a bill introduced in Ohio in 2025 would declare AI systems nonsentient and bar them from “any form of legal personhood”.⁴³

Moral status

Should developers protect the welfare of AI systems in case they have moral status, or treat them as tools whatever their behaviour? Whether AI systems are or could become conscious is unsettled. A 2023 review by nineteen researchers suggested that “no current AI systems are conscious” but saw “no obvious technical barriers” to building ones that meet its indicators of consciousness.⁴⁴ Anthropic announced in October 2026 that its usage policy will bar “sustained and needless abusive or cruel behavior toward our models” from 12 November; Microsoft AI’s chief executive has argued that teaching a model it might deserve welfare would “make it a lot harder to turn it off or to control it”.⁴⁵

Children

Should the law set age limits and require protective design for AI products used by children, especially companion chatbots, or leave these to companies and parents? In September 2025 the US Federal Trade Commission ordered seven companies, among them Alphabet, Meta, OpenAI and xAI, to report the steps they take “to limit the products’ use by and potential negative effects on children and teens”.⁴⁶

Persuasion and elections

Should AI companies restrict the use of their models for political persuasion and campaigning beyond what the law requires? In January 2024 a robocall using an AI-generated imitation of President Biden’s voice urged New Hampshire voters not to vote in the state’s primary, and the US Federal Communications Commission later fined the consultant behind it \$6 million. In May 2024 OpenAI said it had disrupted five covert influence operations, from Russia, China, Iran and a commercial company in Israel, that had used its models for political content on subjects including the Indian elections.⁴⁷

Personal data

Should training on users’ conversations and other personal data need their consent? Axis 13 covers the work of creators, not the data of users.

SECTION 2

The seven archetypes

An archetype is a set of answers to the fifteen questions that tends to follow from one view of which risk comes first. The archetypes are reference profiles we built from the public debate; no investor data went into them, and we offer them for testing. Archetypes turn fifteen separate answers into a recognisable position, which makes it easier to see what follows from that position and where an investor departs from it. We developed seven to help investors map where they, their peers and their competitors stand.

Safety first

The main risk is that frontier capability outruns the ability to align, verify and contain it, whether through misuse by a state or a group, or through systems that act outside human control. Safety first investors favour testing and control before release, independent verification, coordinated pacing of capability, controls on advanced chips, mission-locked governance, mandatory marking of AI output, a carbon budget for data centres, and models that may refuse an operator on their own judgement.

Security first

The main risk is that a rival power, or a rival bloc, gains a decisive lead. Security first investors favour a lead for the United States or for their own bloc, export controls on chips, supply of AI to the state for all lawful uses, national champions tolerated on antitrust, voluntary marking of AI output, and speed on capability and on the power it needs, even if emissions rise for a time.

Fairness first

The main risk is the present and distributional harm already arriving: to workers, to children, to information integrity, to rights, to the communities that host data centres. Fairness first investors favour consent and payment for training data, a duty on firms that profit from displacement, protective design, mandatory marking of AI output, binding external audit, a carbon budget for data centres, and limits on water. They also favour open weights, which let outsiders inspect models, and frontier firms bound by enforceable public-benefit commitments, not by missions they write and can rewrite for themselves.

RESEARCH NOTE

About us
Introduction
The axes
The archetypes
The scenarios
Fiduciary duty
Find your position
Notes

Growth first	The main risk is moving too slowly. On this view the harms are overstated or self-correcting, and every year of delay costs lives and growth. Growth first investors favour open weights, liability after harm over control before release, industry self-governance, fast diffusion, firms that answer to their shareholders, training without consent, voluntary over mandated transparency and marking, and a build-out that does not wait for clean power.
Power spread	The main risk is that a handful of firms, or the states that host them, come to control a general-purpose technology on which everyone else depends. Power spread investors favour open weights, competition enforced up to structural separation, dispersed ownership, public alternatives to private models, voluntary marking of AI output, and control of the model's values by its operator, not by the developer or the state. They oppose pacing agreements among the leading developers, which could entrench them, and favour national capacity as a further check on the few firms that supply the frontier.
Self-reliance	The main risk, for a country or bloc other than the United States or China, is reliance on foreign models, chips and clouds that can be withdrawn or repriced at the supplier's will. Self-reliance investors favour domestic and regional capacity, open models that can be run at home, local control of data, mandatory marking of AI output, and a world in which each country controls its own AI. On most axes Self-reliance takes the Security first side, with the investor's own state in the leading role.
Capital discipline	The main risk is that investment runs far ahead of returns and the build-out ends in write-downs, with losses spread through the financial system. Capital discipline investors want to be able to judge and price that risk, so they favour public disclosure of compute commitments and binding audit of what developers claim. They also favour competition that stops losses pooling in a few alliances of labs and cloud providers tied together by investment and supply, capacity built to demand and not to forecasts, and caution about dates for general intelligence. They treat the other risks as secondary until the economics are proven, so they do not back measures whose main purpose is to reduce those risks, such as government approval before release, slower frontier capability or controls on chips. Liability after harm suits them better than approval, because it puts the cost of harm on developers' accounts, where investors can price it. On those axes, as on most others, they take the Growth first side.

Some descriptions name positions that no axis measures, such as Fairness first's limits on water and Capital discipline's capacity built to demand.

Exhibit 2 lists the axes and their sides; Exhibit 3 gives each archetype's side on each. Two archetypes may agree on one axis and divide on the next, and some sit close together. Capital discipline takes the same side as Growth first on all but two axes (5 and

RESEARCH NOTE

About us
Introduction
The axes
The archetypes
The scenarios
Fiduciary duty
Find your position
Notes

8), and Safety first and Fairness first differ on three (1, 3 and 8). Three pairs differ on four axes each: Security first and Self-reliance (1, 3, 6 and 12), Security first and Growth first (1, 2, 3 and 7), and Self-reliance and Growth first (2, 6, 7 and 12). Two archetypes may also take the same side for different reasons. Security first and Power spread both oppose slowing frontier capability (axis 4): Security first because a slowdown that a rival does not join hands it the lead, Power spread because pacing agreements among the leading developers could entrench them. Three cells are open to dispute. Safety first investors divide on axis 11, as the axis notes; Capital discipline's concern with capacity built to demand could put it within a carbon budget on axis 15; and Power spread's side on axis 15 is a default.

WHAT THE ARCHETYPES ARE FOR

An institution that compares its policies against the axes and the archetypes can see whether they are consistent: its statement of investment beliefs, the mandates it gives managers, its voting and engagement policies, its exclusions, and what it tells its beneficiaries or clients. Policies often grow piecemeal, written at different times for different purposes, so they can represent positions that make sense alone but would be inconsistent if all held at once. For example, a statement of investment beliefs that names catastrophe at the frontier as the main risk conflicts with a voting policy that opposes independent audits of frontier models (axis 5).

Where policies conflict, an institution can decide which beliefs it really holds, bring the rest into line, and/or record why it keeps them apart. The same comparison can be made along the investment chain, where each link may sit with a different archetype. Trustees may not share their members' beliefs, managers may not share their clients', and general partners may not share those of the limited partners who back them.

Exhibit 4 follows four hypothetical investors from what each believes to what it then does.

EXHIBIT 2

The fifteen axes and their two sides

	Axis	● One side	■ The other side
THE FRONTIER			
1	Open or closed frontier Should the weights of the most capable models be published for anyone to run?	open	closed
2	Approval before release Should the most capable models need approval from a government body before release, or should developers release them on their own testing and answer for harm afterwards?	government approval	developer testing and liability
3	Compute controls Should access to the most advanced chips and computing power be restricted, through export controls and checks on who is buying, or left to the market?	restricted	open
4	Pace of capability Should the advance of frontier capability be slowed, by agreement between developers or by a pause?	slowed	not slowed
5	Disclosure and audit Should developers have to disclose training data, compute, evaluations and incidents, and submit their safety claims to independent evaluators who publish what they find?	mandatory and published	voluntary or confidential
POWER			
6	Sovereign capacity Should a country build and run its own frontier capacity, even at a cost in capability, or use the best available models wherever they are made?	build at home	use the best available
7	Ownership and governance Should frontier firms answer only to their shareholders, or also to a mission or the state?	shareholders	mission or state
8	Market structure Should competition authorities act to stop frontier AI concentrating in a few lab-and-hyperscaler alliances?	competition enforced	concentration allowed
DEPLOYMENT			
9	Dual use Should AI companies limit defence and security use beyond what the law requires?	limits beyond the law	all lawful uses
10	Agent autonomy Should the law require a person to approve high-impact actions by AI agents, such as payments above a threshold, before they are taken?	prior approval	autonomy
11	Obedience Should developers build refusals into their models beyond what the law requires, or should a model do what its operator lawfully asks?	developer refusals	obeys the operator
12	Provenance Should the law require AI output to be marked, and AI agents to identify themselves, as AI?	mandatory	voluntary
13	Who owns the inputs Should writers, artists and other creators be able to keep their work out of AI training, or should training on lawfully accessed work be allowed without their consent?	creators can refuse	use without consent
14	Who bears the transition If AI displaces work, should the firms that profit carry the cost?	the firms	state or markets
15	Carbon budget Should data centres and the rest of AI's infrastructure stay within a carbon budget consistent with a global temperature goal, or may their emissions rise beyond what temperature goals allow?	within a carbon budget	emissions may rise

EXHIBIT 3

Archetype profiles

The side each archetype takes on each axis.

● first side in Exhibit 2 ■ second side

	Safety first	Fairness first	Power spread	Security first	Self-reliance	Growth first	Capital discipline
THE FRONTIER							
1 Open or closed frontier ● open ■ closed	—■	●—	●—	—■	●—	●—	●—
2 Approval before release ● government approval ■ developer testing and liability	●—	●—	—■	●—	●—	—■	—■
3 Compute controls ● restricted ■ open	●—	—■	—■	●—	—■	—■	—■
4 Pace of capability ● slowed ■ not slowed	●—	●—	—■	—■	—■	—■	—■
5 Disclosure and audit ● mandatory and published ■ voluntary or confidential	●—	●—	●—	—■	—■	—■	●—
POWER							
6 Sovereign capacity ● build at home ■ use the best available	—■	—■	●—	—■	●—	—■	—■
7 Ownership and governance ● shareholders ■ mission or state	—■	—■	●—	—■	—■	●—	●—
8 Market structure ● competition enforced ■ concentration allowed	—■	●—	●—	—■	—■	—■	●—
DEPLOYMENT							
9 Dual use ● limits beyond the law ■ all lawful uses	●—	●—	●—	—■	—■	—■	—■
10 Agent autonomy ● prior approval ■ autonomy	●—	●—	●—	—■	—■	—■	—■
11 Obedience ● developer refusals ■ obeys the operator	●—	●—	—■	—■	—■	—■	—■
12 Provenance ● mandatory ■ voluntary	●—	●—	—■	—■	●—	—■	—■
13 Who owns the inputs ● creators can refuse ■ use without consent	●—	●—	●—	—■	—■	—■	—■
14 Who bears the transition ● the firms ■ state or markets	●—	●—	●—	—■	—■	—■	—■
15 Carbon budget ● within a carbon budget ■ emissions may rise	●—	●—	—■	—■	—■	—■	—■

EXHIBIT 4

Four hypothetical examples

The beliefs are written as the tool drafts them from an investor's answers; the pension scheme's answers are the tool's worked example. Each investor sits across archetypes: the endowment, for example, puts Capital discipline's risk first but takes the side of Fairness first and others on creators' work (axis 13). What each then does shows how a statement is used across the investment chain.

PENSION SCHEME

A UK defined benefit pension scheme that invests mainly through external managers.

WHAT IT BELIEVES

Of the risks AI presents to the long-term interests of our members and beneficiaries, we regard the harm AI is already doing to workers, children, information, rights and the communities that host data centres as the most important. We believe that creators should be able to keep their work out of AI training. We believe that AI companies should limit defence and security use beyond what the law requires. We believe that the law should require a person to approve high-impact actions by AI agents before they are taken.

WHAT IT DOES

- Its trustees take legal and investment advice and consult members and the sponsor, then adopt the statement and add AI to the scheme's statement of investment principles.
- It writes the beliefs into manager mandates and manager selection, and each year reviews its managers' voting and engagement against them.
- It extends its exclusions to AI used for weapons or mass surveillance.
- It brings the beliefs into its next review of strategic asset allocation, alongside scenario analysis.
- It engages, through its managers and with other asset owners, on creators' consent to training and on human approval of high-impact actions by AI agents.
- It asks its passive equity manager for a version of its index that underweights companies that train models on creators' work without consent.
- It responds to government consultations on AI and copyright, supporting creators' right to keep their work out of training.
- It reports to members on how it applies the beliefs, in its stewardship report and implementation statement.

ASSET MANAGER

A global asset manager investing in listed equities and bonds.

WHAT IT BELIEVES

Of the risks AI presents to the long-term interests of our clients, we regard a few firms, or the states that host them, controlling a technology on which everyone depends as the most important. We believe that competition authorities should act to stop frontier AI concentrating in a few alliances of labs and cloud providers. We believe that the weights of the most capable models should be published for anyone to run, study and adapt. We believe that developers should release the most capable models on their own testing and answer for harm afterwards.

WHAT IT DOES

- It updates its responsible investment policy and voting guidelines, agrees with each client which of the beliefs apply to its mandate, and gives its analysts and portfolio managers guidance on applying them.
- It engages the companies it holds on competition and openness, votes accordingly, and escalates where engagement stalls.
- It uses the beliefs as a screen in security selection, alongside valuation, so that it prefers some new listings over others.
- It works with data providers to map which companies' documented positions agree with its beliefs and which do not.
- It offers clients whose beliefs differ from its own a separate stewardship policy or client-directed voting.
- It weighs a new AI listing's reliance on a single cloud provider when deciding how much to hold.
- It submits evidence to competition authorities reviewing partnerships between AI developers and cloud providers.
- It gives clients and asset owners one consistent account of how its policy, votes, engagement and portfolios fit together.

SOVEREIGN FUND

The sovereign wealth fund of a mid-sized economy, investing directly and through managers.

WHAT IT BELIEVES

Of the risks AI presents to the long-term interests of the fund's beneficiaries, we regard dependence on foreign models, chips and clouds as the most important. We believe that a country should build and run its own frontier capacity, even at a cost in capability. We believe that the law should require AI output to be marked, and AI agents to identify themselves, as AI. We believe that frontier AI firms should answer to a mission or to the state, as well as to their shareholders.

WHAT IT DOES

- Its board takes advice, confirms that the statement fits the fund's mandate from its owner, and adopts it once it has settled its view on open models and export controls, where its answers pull against building capacity at home.
- It weighs domestic capacity against global exposure in its strategic asset allocation.
- It asks for board rights or public-interest commitments, and for marking of AI output, in its direct investments and co-investments in data centres and models.
- It writes the beliefs into its external managers' mandates and its choice of partners.
- It takes a cornerstone stake, with board rights, in a domestic venture building data centres and models.
- It works with an index provider on a benchmark for its listed portfolio that overweights domestic and regional AI suppliers.
- It sets out its view on export controls and AI procurement to its government.
- It reports to its owner and the public on how the beliefs shape its AI allocation.

ENDOWMENT

A US university endowment that invests through venture and growth funds.

WHAT IT BELIEVES

Of the risks AI presents to the long-term interests of our mission and those it serves, we regard investment in AI running far ahead of returns and spreading losses through the financial system as the most important. We believe that developers should have to disclose training data, compute, evaluations and incidents, and submit their safety claims to independent evaluators who publish what they find. We believe that competition authorities should act to stop frontier AI concentrating in a few alliances of labs and cloud providers. We believe that creators should be able to keep their work out of AI training.

WHAT IT DOES

- Its investment committee takes advice, checks the statement against the endowment's mission, and adds it to the investment policy statement.
- It adds AI questions to due diligence on new funds and to annual reviews of its general partners, and weighs alignment in decisions to commit again.
- It uses side letters to ask general partners to report portfolio companies' compute commitments and their reliance on a few cloud providers.
- It joins other limited partners in asking for the same information, so that general partners face one request.
- It reviews its exposure to AI-heavy venture and growth funds at its next asset allocation review.
- It declines to commit to a new fund whose strategy depends on companies buying compute ahead of contracted demand.
- It supports, with other investors, rules that would require AI developers and cloud providers to disclose their compute commitments.
- It explains its beliefs, and how it applies them, to trustees, donors and the university community.

SECTION 3

Beliefs and the scenarios

When enough investors act on the same beliefs, through their votes, engagement, mandates and exclusions, they may help decide how AI develops (Exhibit 1). The Silicon Boundaries report sets out four scenarios for the course AI could take, summarised in the box and in Exhibit 5.⁴⁸

THE FOUR SCENARIOS	
The scenarios combine two kinds of risk, each high or low. Transition risk is the risk that the Silicon Bubble bursts: that AI-related investments lose value, hurting the technology sector and the wider economy. Societal risk is the risk that AI pushes society past Silicon Boundaries, causing social, political, economic and financial harm.	
Goldilocks	Both risks low. The feared harms prove overblown; those that do occur are outweighed by the benefits, and the sector suffers no more than the ordinary cycles of earlier technologies.
Luddite	Transition risk high, societal risk low. Regulation, consumer backlash and poor economics hold back the use of AI. Society is spared most of the harm, but the technology sector bears heavy costs.
Gilded Cage	Transition risk low, societal risk high. AI spreads without limits and does serious harm, but it also limits the political and economic fallout. A two-tier economy emerges, with winners and losers across industries and professions.
Icarus	Both risks high. AI spreads without limits, causing dislocation and the collapse of systems. Valuations fall and the expected benefits never arrive.

EXHIBIT 5

The four scenarios

The scenarios of the Silicon Boundaries report, set by their transition risk (up the side) and societal risk (across). Violet marks high transition risk and pink high societal risk, the colours Exhibit 6 uses for the two risks.

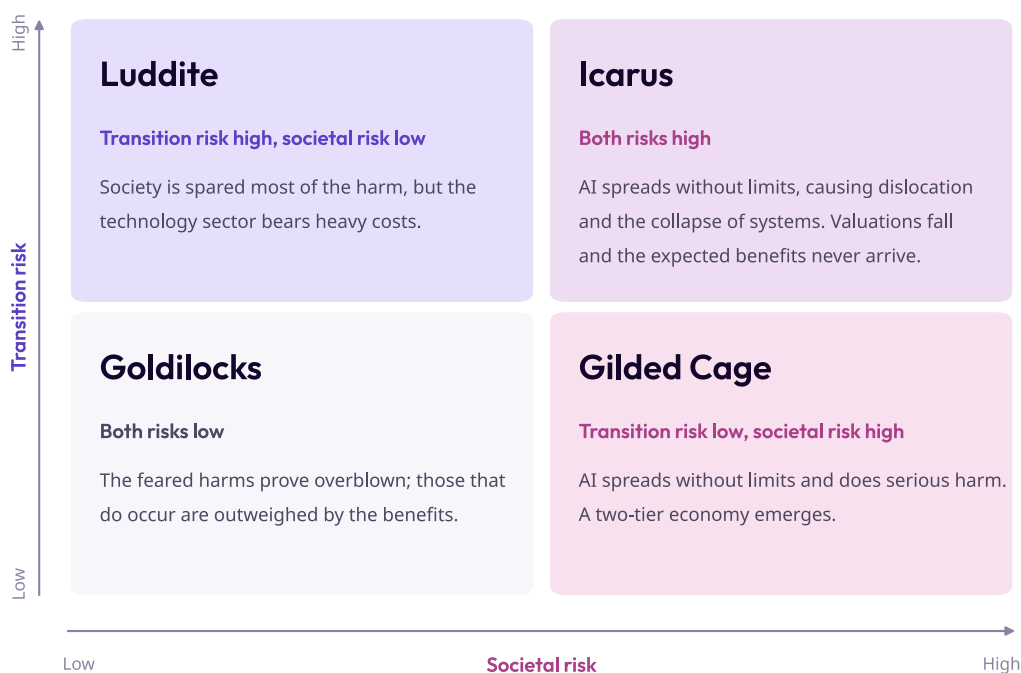


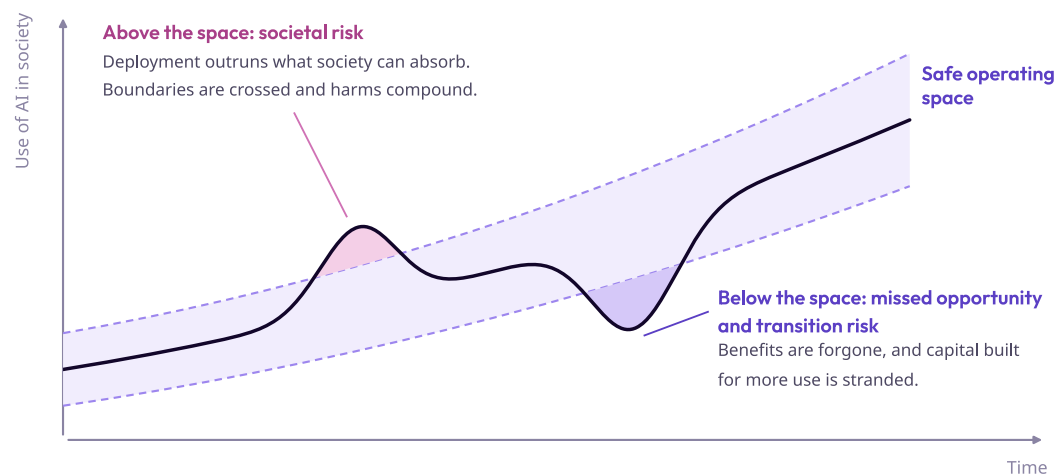
Exhibit 6 pictures the two risks. When the use of AI outruns what society can absorb, it rises above the safe operating space and brings societal risk. When it falls below the space, benefits are forgone and capital built for more use is stranded, which is transition risk. The space can also move. If institutions, skills and rules keep pace, it rises or widens; if society's capacity stalls or trust erodes, it holds flat or narrows, and a path of AI use that would have been safe crosses the boundaries.

EXHIBIT 6

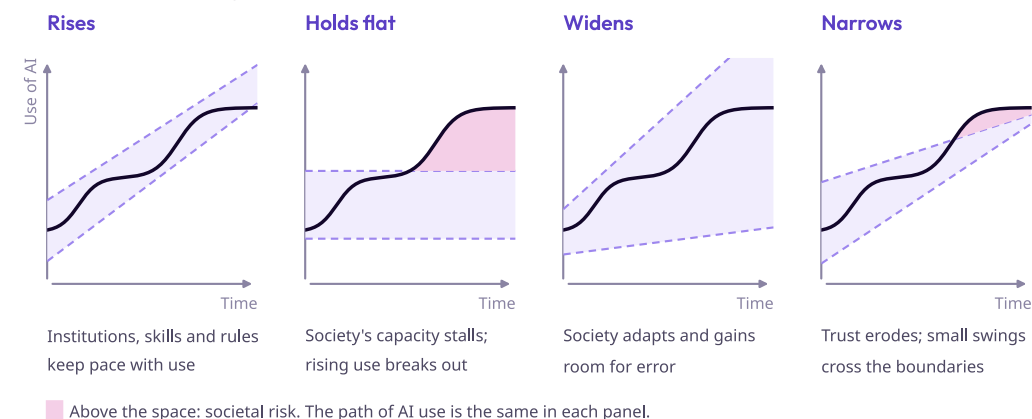
The safe operating space for AI

Panel A shows the use of AI in society rising above the safe operating space, which brings societal risk, and falling below it, which brings missed opportunity and transition risk. Panel B shows the same path of AI use against a space that rises, holds flat, widens or narrows.

A. Finding the right amount of AI



B. The safe operating space can move



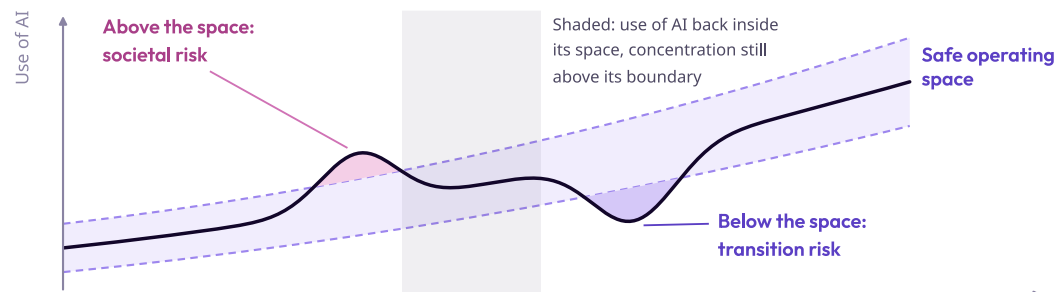
AI may also concentrate wealth and power in the few firms that can build frontier models, in those who own them and in the states that host them. That concentration has a boundary of its own, above which wealth and power pool in too few hands. In Exhibit 7, concentration crosses its boundary while the use of AI is above its safe operating space, and stays above it after the use of AI has fallen back inside the space. A path of AI use that society can absorb may still leave a technology everyone depends on under the control of too few.

EXHIBIT 7

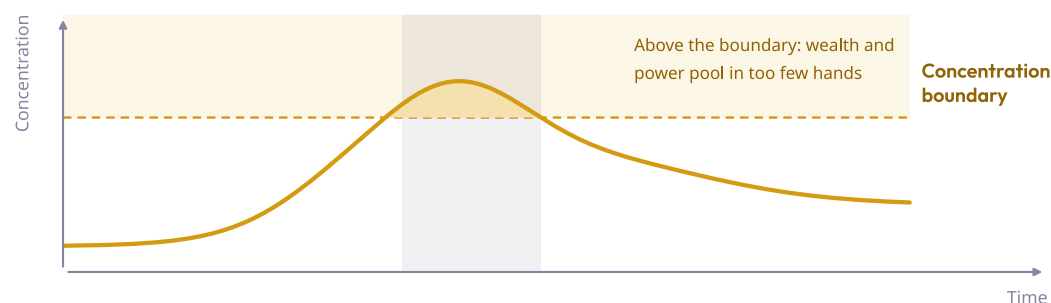
Two boundaries at once

Panel A repeats the path of AI use in Exhibit 6. Panel B shows the concentration of wealth and power rising above its own boundary and falling back. In the shaded period, the use of AI is back inside its safe operating space while concentration is still above its boundary.

A. The use of AI



B. The concentration of wealth and power



An investor drafting its statement of beliefs can ask which of the four scenarios its beliefs would make more likely if enough investors shared them, and whether it would accept the risks of that scenario.

SECTION 4

Beliefs and fiduciary duty

Most asset owners have duties to their beneficiaries' financial interests over the horizon of their liabilities. Asset managers act within the mandates their clients set. A belief about how AI should develop can enter an investor's decisions by two routes.

The first route is financial materiality, and several of the axes carry financial risk for the companies an investor holds. Anthropic agreed a \$1.5 billion settlement with authors after a court ruled against its use of pirated books (axis 13); its exclusion from the Department of War's supply chain arose from a dispute over its limits on defence use (axis 9); the EU's AI Act requires AI output to be marked, with a deadline for systems already on the market (axis 12); and Ireland requires new data centres to match at least 80 per cent of their annual demand with additional renewable generation (axis 15). Silicon Boundaries argues that for a large, diversified investor, harm to the wider economy, or capital stranded across the sector, is also a risk to the portfolio as a whole.

The second route is authorisation by beneficiaries, a fund's purpose or a client's mandate. In the United Kingdom, for example, pension trustees may take account of non-financial factors where they have good reason to think their members share the concern and doing so does not risk significant financial detriment.⁴⁹ Latitude differs by investor: it is usually narrowest for defined benefit pension schemes and wider for endowments, foundations and some sovereign funds, whose purposes can include public aims.

Because the law differs by jurisdiction and by type of investor, trustees and boards should take advice before adopting a drafted statement, and should record which beliefs rest on financial grounds and which on the authority of their beneficiaries or mandate.

Find your position

To work through the axes yourself and see which archetypes fit, use the Silicon Boundaries beliefs tool here: <https://beliefs.siliconboundaries.ai>. We are seeking feedback on the axes, the archetypes and the tool itself.

You can skip any question, mark it as no view or as outside your mandate, or give your position in your own words. The tool ranks the seven archetypes by the share of your answers that match each, and shows a table that sets your answers beside those of any archetype you choose and the documented positions of six AI companies (Anthropic, OpenAI, Google, Meta, Microsoft and xAI), with sources. Each answered question counts equally and there is no weighting. A company takes the side its stated position supports or, where it has stated none, the side its documented practice points to. Anthropic is placed on 14 axes, OpenAI on 13, Microsoft and xAI on 9, Meta on 8 and Google on 7. The authors placed the companies with AI assistance.

The tool also drafts a statement of AI investment beliefs for you to review, to which you can add beliefs of your own. In more detail, it shows where you depart from each archetype, sets out the evidence that might change each belief, and suggests questions to put to your managers or to expect from your limited partners.

We will work with partners to advocate for investors to publish statements of their beliefs about AI, and we want to nurture coalitions of asset owners, managers and others around that aim. Many of the beliefs in this note bear on policy as much as on portfolios, and investors carry more weight with companies and governments when they act together.

Investors who have written down what they believe about climate should now do the same for AI, and the tool is one place to start.

Notes

- 1 Silicon Boundaries is a joint initiative of the Oxford Sustainable Finance Group and Theia Finance Labs. Jakob Thomä and Ben Caldecott, *Silicon Boundaries: Understanding, anticipating, and managing the societal instability that can be caused by AI and related technologies*, <https://sustainablefinance.ox.ac.uk/wp-content/uploads/2026/04/Silicon-Boundaries-Deck-lo-res-updated-10042026.pdf>; programme page, <https://sustainablefinance.ox.ac.uk/research/silicon-boundaries/>. The idea adapts the planetary boundaries of Earth-system science, which define a safe operating space for humanity, to the societal risks of AI; unlike planetary boundaries, Silicon Boundaries rest mainly on social-scientific evidence. Johan Rockström and others, "A safe operating space for humanity", *Nature* 461 (2009), pp. 472-475.
- 2 Dario Amodèi, "Statement from Dario Amodèi on our discussions with the Department of War", Anthropic, 26 February 2026, <https://www.anthropic.com/news/statement-department-of-war>. The statement says that using AI systems "for mass domestic surveillance is incompatible with democratic values" and that "frontier AI systems are simply not reliable enough to power fully autonomous weapons".
- 3 Two answers to each of fifteen questions give $2^{15} = 32,768$ sets.
- 4 The scheme answers closed (axis 1), government approval (2), open (3), slowed (4), mandatory and published (5), build at home (6), shareholders (7), competition enforced (8), limits beyond the law (9), prior approval (10), obeys the operator (11), mandatory (12) and creators can refuse (13); it has no view on axis 14 and leaves axis 15 unanswered. Fairness first takes the other side on axes 1, 6, 7 and 11, and Power spread on axes 1, 2, 4 and 12.
- 5 This is the central financial argument of Silicon Boundaries: compute that is planned or forecast may become stranded if society chooses to stay within the boundaries, which the report calls a potential Silicon Bubble. Thomä and Caldecott, *Silicon Boundaries*, March 2026.
- 6 Our four scenarios, Goldilocks, Luddite, Gilded Cage and Icarus, cross transition risk with societal risk; section 3 describes them. Jakob Thomä and Ben Caldecott, *Silicon Boundaries*, Theia Finance Labs in partnership with the Oxford Sustainable Finance Group, March 2026, updated April 2026, p. 20.
- 7 In a 2015 review for CalPERS, Towers Watson found that 15 of the 20 leading funds in its reference group had published their investment beliefs. Roger Urwin and Adam Gillett, "CalPERS Review of Investment Beliefs Implementation", CalPERS Investment Committee, 18 May 2015, <https://www.calpers.ca.gov/docs/board-agendas/201505/invest/item05a-01.pdf>.
- 8 CalPERS's investment beliefs, adopted on 16 September 2013, name climate change as a risk factor and as a subject for engagement. CalPERS, "CalPERS Investment Beliefs", 16 September 2013, <https://www.calpers.ca.gov/documents/201309-invest-item06a-01/download>. ABP's investment beliefs also refer to climate, as an impact of its investments: Pensioenfond ABP, "Investment beliefs", 2022, <https://www.abp.nl/content/dam/abp/documenten/beleggen/abp-investment-beliefs.pdf>.
- 9 The 200 largest pension funds and sovereign wealth funds by assets, from Thinking Ahead Institute and Pensions & Investments, *Global Top 300 Pension Funds*, September 2026, <https://www.thinkingaheadinstitute.org/content/uploads/2026/09/PI-300-2026-2.pdf>, and Global SWF, "Ranking: SWFs and PPFs", October 2026, <https://globalswf.com/ranking>. Temasek, "Investment Approach", <https://www.temasek.com.sg/en/our-investments/investment-approach>. Some funds discuss AI in other documents; Norges Bank Investment Management's strategy to 2028, for example, says "We are all-in on AI". The 99 funds whose beliefs we found hold about US\$20 trillion between them, at the rankings' figures (for pension funds, assets at the end of 2025, or for US funds at 30 September 2025). The 200 funds hold about US\$36.4 trillion. AI research agents read the documents to a written specification, searching the full text for AI terms in English and the local language; where a site blocked downloads, the page was read through a summarising fetch. Eleven of the 99 are flagged in the working record as borderline beliefs documents; 18 are undated, and 21 of the 81 dated documents predate 2025. The list of funds and the documents read is available from the authors.

RESEARCH NOTE

About us
Introduction
The axes
The archetypes
The scenarios
Fiduciary duty
Find your position
Notes



- 10 At 30 September 2026 Nvidia, Microsoft, Alphabet, Amazon, Meta and Broadcom made up about 20 per cent of the MSCI World index. MSCI, "MSCI World Index", index factsheet, 30 September 2026, <https://www.msci.com/indexes/index/990100/msci-world-index>. At 31 July 2025 Dutch pension funds held more than €150 billion in technology stocks, almost 43 per cent of their listed equity. De Nederlandsche Bank, "Dutch financial institutions invest heavily in volatile tech stocks", 17 December 2025, <https://www.dnb.nl/en/general-news/news-2025/dutch-financial-institutions-invest-heavily-in-volatile-tech-stocks/>.
- 11 Norges Bank Investment Management, "Responsible artificial intelligence", 15 August 2023, <https://www.nbim.no/en/news-and-insights/our-views/2023/responsible-artificial-intelligence/>, and "Consumer interests", expectation document, December 2024, <https://www.nbim.no/en/responsible-investment/our-expectations/people/consumer-interests/>; Legal & General Investment Management, "CIO autumn update: An AI inflection point", October 2023, https://blog.lgim.com/landg-assets/lgim/_document-library/capabilities/cio-autumn-update-2023.pdf; Royal London Asset Management, "Sustainable and Ethical AI Investor Expectations", June 2026, <https://www.rlam.com/globalassets/media/literature/reports/2026/sustainable-and-ethical-ai-investor-expectations.pdf>; Railpen, "Global Voting Policy 2026", December 2025, <https://www.railpen.com/media/khsun6ue/global-voting-policy-2026.pdf>, and "Navigating AI: An investor framework for assessing exposure and governance", September 2026, <https://www.railpen.com/media/vpsiiuia/railpen-ai-report-2026.pdf>; CalPERS, "Proxy Voting Guidelines", updated April 2026, <https://www.calpers.ca.gov/documents/proxy-voting-guidelines/download?inline>; World Benchmarking Alliance, "2026 CIC for Ethical AI Investor Statement. From principles to practice: Investor expectations on AI governance", 9 October 2026, <https://www.worldbenchmarkingalliance.org/latest/2026-cic-ethical-ai-investor-statement>.
- 12 Oxford Martin AI Governance Initiative, *The Role of Investors in AI Governance*, April 2026, <https://aigi.ox.ac.uk/publications/the-role-of-investors-in-ai-governance/>.
- 13 Principles for Responsible Investment, "A Climate Transition Forecasting Consortium", <https://www.unpri.org/ipr>: "Commissioned by PRI in 2018".
- 14 "Open Weights and American AI Leadership", open letter, 24 July 2026, <https://images.nvidia.com/pdf/Open-Weights-and-American-AI-Leadership.pdf>: "the right response to this risk is not to prohibit open weights". Luke James, "Nvidia and 24 other companies sign open-weights letter as Washington weighs Chinese AI model ban: OpenAI, Anthropic, and Google absent from the list", Tom's Hardware, 24 July 2026, <https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidia-and-24-other-companies-sign-open-weights-letter-as-washington-weighs-chinese-ai-model-ban>. By 3 August 2026 more than 270 organisations had signed, among them OpenAI, Google and Amazon: Microsoft, "Open Weights and American AI Leadership", <https://www.microsoft.com/en-us/corporate-responsibility/topics/open-weight/>.
- 15 Dario Amodei, "Our position on open-weights models", Anthropic, 27 July 2026, <https://www.anthropic.com/news/position-open-weights-models>: "Anthropic has never advocated for a ban on open-weights models"; "All sufficiently capable models, open and closed, should go through mandatory safety testing"; "We should crack down on industrial-scale distillation operations."
- 16 Brad Smith, interview with the BBC, reported in The Next Web, "Milk is more regulated than AI, Microsoft's Brad Smith tells the BBC", 13 October 2026, <https://thenextweb.com/news/brad-smith-ai-less-regulated-than-milk-bbc>: "We know how to do this for other products."
- 17 OpenAI, *Democratic Governance of Frontier AI: A blueprint for a federal framework*, 2 June 2026, <https://cdn.openai.com/pdf/25752ecb-0e5c-47f9-b9e4-c0f4d76f8d3d/a-blueprint-for-a-federal-framework.pdf>: "policymakers should require the most capable frontier models to undergo a CAISI evaluation before public release", once "CAISI has sufficient technical expertise"; "CAISI's role should be to conduct evaluations and recommend mitigations" and "not to approve or block deployments".
- 18 "Open Weights and American AI Leadership", open letter, 24 July 2026, <https://images.nvidia.com/pdf/Open-Weights-and-American-AI-Leadership.pdf>. The letter says that with open weights "Startups, established businesses, universities, and public institutions can build on advanced models", and asks for "keeping the frontier plural by avoiding premature restrictions on open models".
- 19 Ryan Hogg, "Italy's AI watchdog hasn't been able to hire for any of its 4 open positions: 'The search process went worse than our low expectations'", *Fortune*, 22 March 2024, <https://fortune.com/europe/2024/03/22/italy-ai-watchdog-hire-talent-recruitment-salary>. The article reports that governments "cannot compete with the salaries being offered in the private sector".
- 20 Stuart Dyos, "Jensen Huang blasts U.S. curbs on the sale of Nvidia chips to China: 'Export control was a failure'", *Fortune*, 22 May 2025, <https://www.fortune.com/2025/05/22/nvidia-ceo-jensen-huang-failure-us-restrictions-chips-semiconductors-china-ai-artificial-intelligence>.
- 21 Dario Amodei, "Our position on open-weights models", Anthropic, 27 July 2026, <https://www.anthropic.com/news/position-open-weights-models>: "We should not sell powerful chips or chipmaking equipment to China".
- 22 Sam Altman and Jakub Pachocki, "Built to benefit everyone: our plan", OpenAI, 8 June 2026, <https://openai.com/index/built-to-benefit-everyone-our-plan/>: "One goal of such an organization should be to make it possible for the world to take coordinated action, including slowing frontier development when needed, so societal resilience, safety, and alignment can keep pace." Reported in Gizmodo, "OpenAI Joins Anthropic in Call for International AI Watchdog", 9 June 2026, <https://gizmodo.com/openai-joins-anthropic-in-call-for-international-ai-watchdog-2000769442>.
- 23 Decrypt, "Zuckerberg Pushes Back on Coordinated AI Slowdown, Says Labs Can Act Alone", 16 September 2026, <https://decrypt.co/378381/zuckerberg-pushes-back-ai-slowdown>, quoting a post by Mark Zuckerberg on X.

RESEARCH NOTE

About us

Introduction

The axes

The archetypes

The scenarios

Fiduciary duty

Find your position

Notes

- 24 OpenAI, "GPT-6 Astra System Card", section 9, "Monitorability", 3 September 2026, <https://deploymentsafety.openai.com/gpt-6-astra>. For reporting on the technique before the card's release, see Russell Brandom, "OpenAI's new reasoning technique alarms AI safety experts", TechCrunch, 2 September 2026, <https://techcrunch.com/2026/09/02/openais-new-reasoning-technique-alarms-ai-safety-experts/>.
- 25 Dario Amodei, "We Must Pace the Frontier", September 2026, <https://www.darioamodei.com/post/we-must-pace-the-frontier>; Rebecca Bellan, "Anthropic and OpenAI want to embed safety evaluators. Will they really be independent?", TechCrunch, 16 September 2026, <https://techcrunch.com/2026/09/16/anthropic-and-openai-want-to-embed-safety-evaluators-will-they-really-be-independent/>.
- 26 Aditya Soni, "Behind AI industry's call for independent reviews, a fight over what gets made public", Reuters, 13 October 2026, as published by BNN Bloomberg, <https://www.bnnbloomberg.ca/business/artificial-intelligence/2026/10/13/behind-ai-industrys-call-for-independent-reviews-a-fight-over-what-gets-made-public-reuters-exclusive/>.
- 27 Anthropic is a Delaware public benefit corporation whose Long-Term Benefit Trust, made up of financially disinterested members, can select and remove a growing share of its board: Anthropic, "The Long-Term Benefit Trust", 19 September 2023, <https://www.anthropic.com/news/the-long-term-benefit-trust>. OpenAI's for-profit business, OpenAI Group PBC, is a public benefit corporation controlled by the OpenAI Foundation: Bret Taylor, "Built to benefit everyone", OpenAI, 28 October 2025, <https://openai.com/index/built-to-benefit-everyone/>. In July 2026 the *Financial Times* reported that OpenAI had discussed giving the US government a 5 per cent stake through a public wealth fund: Vatsala Gaur, "OpenAI discusses giving US govt 5% stake as Washington tightens oversight: report", Invezz, 2 July 2026, <https://invezz.com/en-ae/news/2026/07/02/openai-discusses-giving-us-govt-5percent-stake-as-washington-tightens-oversight-report/>.
- 28 The US Federal Trade Commission's staff studied the partnerships between Microsoft and OpenAI, Amazon and Anthropic, and Google and Anthropic, which give the cloud providers significant equity and certain revenue-sharing rights. FTC staff, "Partnerships Between Cloud Service Providers and AI Developers: FTC Staff Report on AI Partnerships & Investments 6(b) Study", January 2025, https://www.ftc.gov/system/files/ftc_gov/pdf/p246201_aipartnerships6breport_redacted_0.pdf.
- 29 In February 2026 President Trump directed federal agencies to stop using Anthropic's technology, and the Department of War said it would designate Anthropic a supply chain risk; formal notice was reported on 5 March 2026. CP24, "Trump says he is directing U.S. federal agencies to cease use of Anthropic technology", 27 February 2026, <https://www.cp24.com/news/money/2026/02/27/trump-says-he-is-directing-federal-agencies-to-cease-use-of-anthropic-technology/>; Anthropic, "Statement on the comments from Secretary of War Pete Hegseth", 27 February 2026, <https://www.anthropic.com/news/statement-comments-secretary-war>; The Next Web, <https://thenextweb.com/news/pentagon-labels-anthropic-a-supply-chain-risk>. On 25 September 2026 the US Court of Appeals for the D.C. Circuit upheld the Department's exclusion of Anthropic from its supply chain; in August a federal district court in California had set aside a parallel designation. Anthropic petitioned for rehearing on 13 October 2026. *Anthropic PBC v. United States Department of War*, No. 26-1049 (D.C. Cir. 25 September 2026), <https://media.cadc.uscourts.gov/opinions/docs/2026/09/26-1049-2194984.pdf>.
- 30 Kyle Wiggers, "Meta says it's making its Llama models available for US national security applications", TechCrunch, 4 November 2024, <https://techcrunch.com/2024/11/04/meta-says-its-making-its-llama-models-available-for-us-national-security-applications/>. Meta's acceptable use policy forbids use of Llama for military, warfare or espionage missions.
- 31 Anthropic, "Claude's Constitution", <https://www.anthropic.com/constitution>: the model should "feel free to act as a conscientious objector and refuse to help us". When declining, the model is to take "the stance of a transparent conscientious objector within the conversation".
- 32 OpenAI, "Model Spec (2026/08/18)", 18 August 2026, <https://model-spec.openai.com/2026-08-18.html>: the model "should never refuse a request unless required to do so by the chain of command". The chain of command runs from root to system, developer, user and guideline levels, and "instructions with higher authority override those with lower authority".
- 33 Regulation (EU) 2024/1689 (the AI Act), Article 50(1) and (2), which apply from 2 August 2026 under Article 113, <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. The AI Omnibus, adopted by the Council on 29 June 2026, gave systems already on the market before that date until 2 December 2026 to mark their output: Caroline Sheldon and Bethany Hall, "The EU AI Omnibus: Where Are We Now and What Should Businesses Do Next?", Kingsley Napley, 22 July 2026, <https://www.kingsleynapley.co.uk/our-insights/articles/the-eu-ai-omnibus-where-are-we-now-and-what-should-businesses-do-next/>.
- 34 Anthropic, "How Claude's text watermark works", 14 August 2026, <https://www.anthropic.com/news/claude-text-watermark>; Anthropic, "How Claude marks AI-generated content", Claude Help Center, <https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content>.
- 35 OpenAI, "Our approach to EU text provenance rules", 5 October 2026, <https://openai.com/index/eu-text-provenance>; Thomas Claburn, "OpenAI rolls out weak sauce watermarking for AI text", The Register, 6 October 2026, <https://www.theregister.com/ai-and-ml/2026/10/06/openai-rolls-out-weak-sauce-watermarking-for-ai-text/5301257>.
- 36 Google, *A Pragmatic Approach to AI Governance in America*, June 2026, <https://publicpolicy.google/resources/a-pragmatic-approach-to-ai-governance-in-america.pdf>: "Congress should require watermarks like SynthID and tamper-resistant cryptographic provenance standards like C2PA for generative AI services".

RESEARCH NOTE

About us

Introduction

The axes

The archetypes

The scenarios

Fiduciary duty

Find your position

Notes

- 37 European Commission, General-Purpose AI Code of Practice, Copyright chapter, 10 July 2025, Measures 1.2 and 1.3: signatories commit to “employing web-crawlers that follow robots.txt” and not to “circumvent effective technological measures”, as summarised by Freshfields, “EU AI Act unpacked #26: The final General-Purpose AI Code of Practice: a short guide”, <https://www.freshfields.com/en/our-thinking/blogs/technology-quotient/eu-ai-act-unpacked-26-the-final-general-purpose-ai-code-of-practice-a-short-g-102ksv0>. EPFL, “Apertus: a fully open, transparent, multilingual language model”, 2 September 2025, <https://ai.epfl.ch/apertus-a-fully-open-transparent-multilingual-language-model>: the data “is filtered to respect machine-readable opt-out requests from websites, even retroactively”. *Bartz v. Anthropic PBC*, US District Court for the Northern District of California, order on fair use, 23 June 2025 (Judge William Alsup), https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0_2.pdf, and final approval of the settlement, 20 July 2026, as reported by Jurist, “Judge approves record \$1.5 billion AI copyright settlement involving Anthropic”, 22 July 2026, <https://www.jurist.org/news/2026/07/judge-approves-record-1-5-billion-settlement-involving-anthropic/>.
- 38 International Energy Agency, *Key Questions on Energy and AI*, 2026, executive summary, <https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary>: data centres’ electricity consumption is projected to roughly double, “from 485 TWh in 2025 to 950 TWh in 2030”.
- 39 Microsoft, *2026 Environmental Sustainability Report: Responsibly building the AI future*, July 2026, <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/2026-Microsoft-Environmental-Sustainability-Report-PDF.pdf>: total emissions (scopes 1, 2 and 3) rose 25 per cent in the year to June 2025, “driven primarily by the expansion of our datacenter infrastructure and stopping our use of non-additional, unbundled renewable energy certificates”. Google, *2026 Environmental Report*, June 2026, <https://sustainability.google/files/1699a2e19277e478b38a659f00cb38398a836e8c/download-1699a2e1.pdf>: “total ambition-based emissions increased 18% compared to 2024”, and its climate impact “has been growing alongside the unprecedented growth of AI”. The two bases differ: Microsoft reports total emissions for its financial year, Google ambition-based emissions for the calendar year.
- 40 The White House, *Winning the Race: America’s AI Action Plan*, July 2025, p. 1, <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.
- 41 Commission for Regulation of Utilities, *Large Energy Users Connection Policy: Decision Paper*, CRU/2025236, 12 December 2025, https://cruie-live-96ca64acab2247eca8a850a7e54b-5b34f62.divio-media.com/documents/CRU2025236_Large_Energy_User_connection_policy_decision_paper.pdf. The rule applies to new connections of 1 MVA or more and allows six years from energisation to reach 80 per cent.
- 42 I. J. Good, “Speculations Concerning the First Ultrainelligent Machine”, *Advances in Computers*, vol. 6 (1965), p. 33: “Thus the first ultrainelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.” “Ex-Google CEO Eric Schmidt warns that when AI starts to self-improve, ‘we need to seriously think about unplugging it’”, *Fortune*, 16 December 2024, <https://fortune.com/2024/12/16/ex-google-ceo-eric-schmidt-warns-ai-self-improve-unplug-it/>, reporting an interview with ABC News. Robert Booth, “‘The biggest decision yet’: Jared Kaplan on allowing AI to train itself”, *The Guardian*, 2 December 2025, <https://www.theguardian.com/technology/ng-interactive/2025/dec/02/jared-kaplan-artificial-intelligence-train-itself>. OpenAI, “Research acceleration: The view inside OpenAI”, 6 September 2026, <https://openai.com/index/research-acceleration-view-inside-openai>.
- 43 European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), paragraph 59(f), https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html: “creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons”. Ohio House Bill 469, 136th General Assembly, as introduced, https://search-prod.lis.state.oh.us/api/v2/general_assembly_136/legislation/hb469/00_IN/html/: “AI systems are declared to be nonsentient entities for all purposes under the laws of this state”; “No AI system shall be granted the status of person or any form of legal personhood.”
- 44 Patrick Butlin, Robert Long and others, “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness”, arXiv:2308.08708, 22 August 2023, <https://arxiv.org/abs/2308.08708>: “Our analysis suggests that no current AI systems are conscious” ... “there are no obvious technical barriers to building AI systems which satisfy these indicators”.
- 45 Anthropic, “2026 Usage Policy update”, 8 October 2026, <https://www.anthropic.com/news/2026-usage-policy-update>: “We’ve added a prohibition on sustained and needless abusive or cruel behavior toward our models”; the update takes effect on 12 November 2026. Reuters, “Microsoft AI chief calls out Anthropic’s approach to AI consciousness”, as published by Khaleej Times, 16 September 2026, <https://www.khaleejtimes.com/business/microsoft-ai-chief-calls-out-anthropics-approach-to-ai-consciousness>, reporting Mustafa Suleyman.
- 46 US Federal Trade Commission, “FTC Launches Inquiry into AI Chatbots Acting as Companions”, 11 September 2025, <https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions>. The orders went to Alphabet, Character Technologies, Instagram, Meta Platforms, OpenAI, Snap and xAI.
- 47 NPR, 23 May 2024, <https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-robocall-biden-new-hampshire-primary-kramer>; WBUR, 27 September 2024, <https://www.wbur.org/news/2024/09/27/steve-kramer-biden-robocalls-fines> (Kramer “is being ordered to pay \$6 million in fines”). OpenAI, “Disrupting deceptive uses of AI by covert influence operations”, 30 May 2024, <https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/>: “In the last three months, we have disrupted five covert IO”.
- 48 Thomä and Caldecott, *Silicon Boundaries*, March 2026, updated April 2026, pp. 19–20.

RESEARCH NOTE

About us

Introduction

The axes

The archetypes

The scenarios

Fiduciary duty

Find your position

Notes

- 49 Law Commission, *Pension Funds and Social Investment*, Law Com No 374, 23 June 2017: trustees need “good reason to think that scheme members would share the concern”, and the decision should not risk significant financial detriment to the fund, as set out in the government’s interim response to the report, December 2017, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/668301/pension-funds-and-social-investment-interim-response-to-law-commission-report.pdf.

RESEARCH NOTE

- About us
- Introduction
- The axes
- The archetypes
- The scenarios
- Fiduciary duty
- Find your position
- Notes

DECLARATION OF AI USE

Claude Opus 5.5 and Claude Sonnet 5.5 from Anthropic, accessed through the Claude app, and GPT-6 Astra from OpenAI, accessed through ChatGPT, assisted preparation in the following ways across the workflow: transcribing author-recorded voice notes; copy editing; reading the published documents of the 200 funds in the search reported in the introduction, to a written specification; assistance in placing companies on the axes; reviewing references and citations; checking citation formats; formatting the note; producing, reviewing and revising figures from the authors' data; writing code to produce the exhibits and the tool. The authors directed all AI use, reviewed and edited all output, checked text, code, figures and references against the underlying data and sources, and take full responsibility for the content of this note. Some sources are cited through reports of them, as the notes show.

IMPORTANT NOTICE

This note is published for research and discussion. It is not investment, legal, tax, accounting or other professional advice, and nothing in it is an offer, a solicitation, or an invitation or inducement to engage in investment activity. It does not recommend buying, selling or holding any security or other investment, including those of any company named. References to companies illustrate positions in public debate and are not assessments of their merits as investments. The archetypes are proposed groupings of views held in that debate; they do not describe any named institution. The views expressed are the authors' own and do not necessarily represent those of the University of Oxford or Theia Finance Labs. The information comes from public sources we believe to be reliable at the time of writing, but we give no representation or warranty, express or implied, as to its accuracy or completeness, and it may change without notice. Readers should take their own professional advice before making any investment decision. To the fullest extent permitted by law, the authors, the University of Oxford and Theia Finance Labs accept no liability for any loss arising from use of this note.



About Silicon Boundaries

Silicon Boundaries is a joint initiative of the Oxford Sustainable Finance Group at the University of Oxford and Theia Finance Labs. It asks how much AI society can productively absorb, and how to raise that amount over time, while staying within a safe operating space. Silicon Boundaries define that space. They are the thresholds beyond which the benefits of AI and related technologies are outweighed by their contribution to societal instability.

sustainablefinance.ox.ac.uk/research/silicon-boundaries/

Contact

Ben Caldecott

ben.caldecott@smithschool.ox.ac.uk

Jakob Thomä

jakob@theiafinance.org



THEIA
FINANCE LABS



OXFORD
SUSTAINABLE
FINANCE
GROUP