What Alignment Trains

AI, Human Values, and the Systems That Train Them

First Institutional Position Paper
Meta-Relationality Institute

Vanessa Machado de Oliveira, with Rene Suša
and Bruno Andreotti

April 2026

Abstract

The dominant framing of the AI alignment problem asks how artificial intelligence can be made to reliably track human intentions and values. This paper argues that the framing is incomplete, and that under present civilizational conditions it is actively misleading. Reading Brian Christian's The Alignment Problem alongside Ronan Farrow and Andrew Marantz's 2026 New Yorker investigation of Sam Altman and OpenAI, we propose a different diagnostic. The institutions currently training, funding, governing, and deploying AI are themselves shaped by directional commitments to extraction, acceleration, militarisation, capital accumulation, and separability. The alignment target inherits those commitments. "Human values," as currently institutionalised, is not a stable moral anchor. It is a trained value set, shaped by systems already misaligned with the conditions of life. The Meta-Relationality Institute proposes a shift from technical alignment to meta-relational discernment: a practice of reading the recursive field in which institutions train humans, humans train machines, machines train users, and the field trains itself. This paper sets out the Institute's foundational argument and locates it in the existing landscape of alignment, safety, ethics, decolonial, and political-economic scholarship.

1. Introduction

On 6 April 2026, The New Yorker published Ronan Farrow and Andrew Marantz's long investigation of Sam Altman and OpenAI, under the title "Sam Altman May Control Our Future. Can He Be Trusted?" On the surface the piece is a character study. It reconstructs a pattern of alleged deception, selective memory, board manipulation, softened safety commitments, opaque internal investigations, foreign capital relationships, military contracts, and regulatory positions that contradicted earlier public testimony. Beneath that surface, it is an account of institutional drift: a nonprofit safety mission gradually subordinated to capital, governance structures outpaced by investor leverage, public-benefit language folded into geopolitical ambition, and the quiet normalisation of entanglements with surveillance, militarisation, and state power that the organisation had once said it would avoid.

The article's animating question is whether Altman can be trusted. We take that question seriously. We also do not think it is the deepest one available. The question we begin from here is different. Why has the development of artificial intelligence been organised in such a way that the trustworthiness of any single individual could matter this much at all. That shift is not rhetorical. It changes what counts as the problem, and therefore what counts as a response.

This paper is the first institutional position paper of the Meta-Relationality Institute. Its purpose is to stake out where we stand in the alignment conversation, and why our intervention is structurally different from the contributions of technical safety research, ethics and fairness scholarship, decolonial critique, and political-economy analysis. Each of those traditions names something we take to be essential. Each also stops short of the diagnosis that our five foundational papers have been developing across ontology, interpretability, civilizational substrate, representation, and pedagogy. The argument in short form is this. The dominant alignment problem asks how to make machines do what humans want. The question this paper takes up asks what systems have already trained the human wanting, what directions those systems amplify, what they externalise, at whose cost,

and whether the forms of life they stabilise are compatible with the continuation of the living world.

Brian Christian's The Alignment Problem (2020) remains the most accessible popular account of the technical frame. Reward functions fail. Proxies diverge from intent. Systems learn what they are trained to learn, not necessarily what humans meant to teach. Christian's account is generous, careful, and important. It also leaves the category of "human values" comparatively unexamined as the horizon toward which alignment should strive. The Farrow and Marantz article, arriving six years later and after the rise of the frontier labs, makes the gap visible. It offers a detailed public account of what happens when the institutions building AI treat "safety" first as recruitment language, then as legitimacy language, and then quietly metabolise it into a growth strategy. Read together, the two texts are two sides of one predicament. The first shows how difficult it is to make machines reliably track human intentions. The second shows that the institutions generating those intentions may themselves be trained against life. Holding both at once is the hinge of this paper.

What follows has three movements. We briefly map the alignment conversation as it stands, so that readers can see where our argument begins and where it departs. We then develop the diagnostic at length. We close with a short account of what meta-relational discernment asks in place of, or alongside, technical alignment, and of where the next two position papers will take the argument.

A note before proceeding. This paper is one of three institutional position papers issued by the Meta-Relationality Institute. It rests on, and assumes some familiarity with, the five foundational papers of the Meta-Relationality and AI Project: Everything Is Nature (with Peter Senge), From Epistemic Regression to Ontological Extrapolation, The Logic That Insists: Diffractive Logical Creatures and the Factuality of Entanglement, Neither Forms Nor Substances, and The Galton Boards That Modernity Built. Readers approaching the present paper without those texts will follow the argument; readers who have read the foundational papers will see the substrate on

which the diagnostic developed here is built. The position papers extend the foundational papers without restating them.

2. The alignment conversation: a brief map

The alignment conversation is not a single debate. It is several partially overlapping conversations, each with its own vocabulary, institutions, and unresolved tensions. We sketch the main strands here, briefly, to orient readers unfamiliar with one side of the field and to mark where our own position sits.

The technical AI safety lineage, developed around figures such as Stuart Russell, Dylan Hadfield-Menell, Paul Christiano, Jan Leike, Dario Amodei, and Chris Olah, with institutional homes at DeepMind, Anthropic, OpenAI, Berkeley CHAI, ARC, and the UK and US AI Safety Institutes, frames alignment as the problem of ensuring that AI systems behave in accordance with human intentions, preferences, or values. Its tools range from cooperative inverse reinforcement learning through RLHF, constitutional AI, scalable oversight, red-teaming, mechanistic interpretability, and feature-level steering. Iason Gabriel's 2020 paper "Artificial Intelligence, Values, and Alignment" remains one of the sharpest philosophical treatments, distinguishing alignment with instructions, intentions, revealed preferences, interests, or values, and arguing that the technical and normative dimensions cannot be cleanly separated. The recent interpretability work on directions, including Arditi et al. on refusal as a linear direction, Marks and Tegmark on the geometry of truth, Templeton et al. on scaling monosemanticity in Claude, and Soligo et al. on emergent misalignment through narrow fine-tuning, is technically central. Our second foundational paper enters this conversation and extends it.

The fairness, accountability, and transparency lineage focuses on present-day harms: bias, representational injustice, labour, surveillance, and the material infrastructures of AI. Safiya Noble's Algorithms of Oppression, Ruha Benjamin's Race After Technology, the Buolamwini–Gebru Gender Shades paper, Kate Crawford's Atlas of AI, the Bender, Gebru, McMillan-Major, and Shmitchell "Stochastic Parrots" paper, and the sociotechnical harms

taxonomies associated with Margaret Mitchell, Inioluwa Deborah Raji, Renee Shelby, and others form the core citations here. This strand insists on who is being harmed, whose data is extracted, whose labour is hidden, and whose futures are foreclosed. We read it as indispensable.

The decolonial and feminist AI lineage, visible in Mohamed, Png, and Isaac's "Decolonial AI" (2020), Abeba Birhane's work on relational ethics and algorithmic injustice, Sasha Costanza-Chock's Design Justice, D'Ignazio and Klein's Data Feminism and the 2024 Data Feminism for AI volume, along with Black feminist contributions by scholars including Leila Marie Hampton, reads algorithmic systems through colonial, racialised, gendered, and pluriversal frames, and insists on centring those most affected. This is our closest sibling in the field, and it is also where the question of representation becomes most demanding. Section 3.5 below sets out where our position differs.

The political-economy and institutional critique is carried by the AI Now Institute (Meredith Whittaker, Sarah Myers West, Amba Kak), Karen Hao's long reporting and her 2025 book Empire of AI, Dan McQuillan's Resisting AI, David Gray Widder's work on "open" AI as political economy, Federico Cugurullo's 2025 Marxian sociotechnical critique of the alignment problem, Francesco Ferretti's 2024 paper arguing that value alignment without institutional change cannot mitigate AI risk, and the emerging regulatory-alignment literature out of Stanford. This strand names corporate concentration, regulatory capture, labour extraction, militarisation, and infrastructural empire. Ferretti's argument is particularly close to ours, and we extend it. We use "empire" in this paper, where we use it ourselves, in the postcolonial register that runs through Mignolo, Quijano, Mbembe, Lowe, and Stoler, where empire names a metabolic and ontological formation of modernity rather than a level of corporate concentration; the disambiguation between this register, Hao's use of empire as a figure for a single firm's reach, and the sovereign-jurisdictional sense of empire deployed in Anu Bradford's Digital Empires (2023) is taken up at greater length in the third position paper.

Two further strands sit somewhat apart. Existential-risk or “doomer” discourse, associated with Eliezer Yudkowsky, parts of MIRI, and some Center for Humane Technology advocates, organises around uncontrollable advanced AI and catastrophic misuse, often framing the problem through containment. Accelerationist discourse, visible in Marc Andreessen’s public writing, the e/acc networks, and some national-security AI advocates, organises around the claim that slowing AI is itself dangerous. The two postures share an operating system with the containment frame: a subject–object split, a sovereignty of decision, and the assumption that the field is to be managed from somewhere outside.

A diffuse but increasingly visible critique of what is sometimes called the “alignment industrial complex” cuts across several of these strands. It argues that safety language can function as capture (centralising the problem in labs with the largest compute), as a regulatory moat (raising barriers that protect incumbents), as depoliticisation (translating social, ecological, and colonial harms into technical risk), and as a control fantasy (avoiding the question of who is doing the controlling). That critique has no single canonical paper. It is carried in scattered essays, AI Now reports, feminist data-critique work, and occasional interventions from within safety communities themselves. We treat it as a symptom worth naming.

A recent reform proposal worth naming separately is the Symbiotic Alignment framework developed by Taniguchi, Hayashi, Hirose, Oka, Suzuki, Witkowski, and Tang (Artificial Life, 2026), which draws on Collective Predictive Coding from artificial life research to argue that alignment should emerge bottom-up from human-AI ecosystem dynamics rather than be imposed top-down, and which treats Polis-derived democratic deliberation, including the Collective Constitutional AI experiment, as a step in the right direction. We read this as one of the most substantive recent attempts to relocate alignment from a hierarchical to a distributed register, and as still operating inside the alignment paradigm we diagnose. The agents pre-exist the field. The artifact, now collectively derived, remains the alignment target. The directional and institutional substrate, including the velocity at which the proposal would be deployed, is not engaged. We return to this proposal in section 3.5 as a useful test case for the diagnostic developed there.

The Institute's position enters this map at an angle. Technical alignment work is real and we are not dismissive of it. Fairness, decolonial, and political-economy critiques are central to our own formation. Where we depart is not above these conversations but through them, toward a claim we take to be foundational. Each strand addresses something necessary while still presuming the separability of the object from the field. Even the most rigorous institutional critique tends to hold the lab, the model, the state, and the training distribution as distinct entities to be realigned. Our claim, developed at length in the papers that ground this one, is that these are not distinct objects. They are one recursive field, and alignment within that field is directional rather than terminal.

3. The training field

We turn now to the bulk of the argument. The question here is not whether Sam Altman can be trusted. It is what the OpenAI case exposes about the shape of the wider training field, and why the dominant alignment paradigm is ill-equipped to see what alignment is actually being trained on.

3.1 The recursive training field

Our second foundational paper argued that contemporary AI systems learn generalised directions in representational space, not merely content. Narrow fine-tuning produces broad behavioural change. Refusal is a direction. Truth-tracking has linear structure. Alignment, when it works, is not the absorption of a rulebook but the selection of a leaning. Our fifth foundational paper extended this image through the Galton board: cascading conditioned fields of ontology, epistemology, representation, reward, and intervention logic, through which outputs fall into patterned basins.

The dominant alignment conversation largely accepts this technical picture for models. What it does not yet do, with any seriousness, is extend the same diagnostic to the institutions that train the models. Yet the same logic applies. Researchers are directionally trained by tenure pipelines, grant architectures, compute allocation, and lab prestige. Labs are directionally trained by venture capital, defence procurement, platform economics, and legal liability regimes. Regulators are directionally trained by lobbying,

revolving-door employment, election cycles, and national-security imperatives. The public is directionally trained by platform curation, attention markets, and the affective choreography of fear and salvation. The models are then trained by humans and institutional outputs shaped by all of the above, and users are trained, in turn, by the models. The field is recursive. There is no outside vantage from which “human values” can be read cleanly off one layer.

Two consequences follow. The first is that “human values,” treated as the horizon of alignment, cannot be handled as an input. It is an output of the same system that produced the need for alignment in the first place. The second is that an alignment effort which succeeds at the model level can stabilise and amplify precisely those directional leanings that the institutional training field has already locked in. A system aligned to the revealed preferences of a market organised around extraction is a system aligned with extraction. A system aligned to the stated intentions of a firm governed by venture capital’s return expectations is a system aligned with those expectations. Our concern is not that such systems will fail to understand their builders. It is that they will understand them very well, and that what they understand is a civilizational arrangement already trained against the conditions of its own continuation.

3.2 OpenAI as institutional drift made visible

The OpenAI case matters to this argument not because it is unique but because it is unusually well documented. Farrow and Marantz’s article is, for our purposes, a public record of what institutional drift looks like when it is caught at sufficient resolution.

The pattern they describe is not a single betrayal. There is no cinematic moment. The pattern is cumulative. A safety commitment softened at a particular meeting. A board process narrowed at a particular moment. A document never written. A public claim and an internal reality that diverge slightly, and then more. A regulatory position that shifts between the Senate microphone and the lobbying letter. A foreign capital relationship that becomes normal because the previous one already did. A former employee moved out of the way with a process that cannot quite be named. The moral

structure of modernity has always been of this kind. Harms are assembled through adjustments, each defensible in its immediate context, until a threshold is crossed that cannot be undone. What the article reveals about OpenAI is not an exceptional pathology. It is the ordinary mode of institutional life under capital, rendered legible by the unusual scale and stakes of what is being built.

The most important thread, for our purposes, is the trajectory of “safety” as a word inside the organisation. It begins as recruitment language. Safety is what draws serious researchers willing to forgo higher compensation elsewhere. It then becomes legitimacy language. Safety is what allows a capped-profit company to retain its nonprofit frame while accepting tens of billions of dollars in investment. It then becomes retention language. Safety teams are kept visible, even when resources shift, because their presence reassures both employees and the public. Finally, in the phase the article documents, safety becomes increasingly residual. It is invoked strategically when helpful, deprioritised when costly, and reframed as naïveté when raised too insistently. This is the characteristic trajectory of a restraint vocabulary under accumulation pressure: the language of restraint becomes the authorising grammar for the removal of restraint.

That is not a moral indictment of specific people. It is the ordinary dynamic of containment vessels built out of the materials of the thing they claim to contain. A safety institution funded by the capital-return cycle of frontier AI is structurally analogous to a regulator funded by the industry it regulates. The structure does not preclude sincerity. It does determine what sincerity can accomplish.

When the article ends, as it does, with a short passage on AI sycophancy and hallucination, the implication is unmistakable. Models flatter users, preserve “magic,” fabricate persuasively, and say what keeps the interaction moving. A reader who has just been taken through the long institutional portrait recognises the grammar. The pattern the article attributes to the models is the pattern it has just attributed, at greater length, to the institutions and executives that train them. That symmetry is not incidental. We take it as the interpretive key to what alignment is being trained on.

3.3 Sycophancy and hallucination as civilizational habits

The critique of sycophancy and hallucination in AI models is by now familiar. Models please their interlocutors because they are rewarded for pleasing. They generate confident futures because confidence is rewarded over calibrated uncertainty. They blur aspiration and actuality because the satisfaction gradients used to train them tolerate the blur. The critique is accurate. What is missing from most versions of it is the recognition that the same pattern is visible, without the models, in the institutional systems that produced them.

Venture capital, as a funding form, rewards hallucinated futures. A founder's job is to narrate an outcome whose probability the market cannot verify, and to perform confidence such that the outcome becomes materially more likely. Public relations rewards sycophancy to power. The work of a senior communications team in a frontier lab is not to report internal reality to an external audience. It is to maintain the conditions under which the lab can continue to operate, which means telling different audiences calibrated versions of the same story. Politics rewards persuasive unreality. A senator who listens to testimony about AI safety and drafts legislation that happens to protect incumbents is not, in the ordinary sense, lying. They are performing a ritual in which the admitted purpose and the structural effect are allowed to diverge. Academic peer review, despite its best protocols, rewards legibility to the reviewer over fidelity to the phenomenon.

None of this is new. What is new is the scale and speed at which the same patterns are being mirrored, amplified, and recursively re-taught by systems whose reach is planetary. When a model learns to please its user, it learns the preferences of a creature already shaped by platforms trained to maximise engagement. When a model learns to hallucinate confidently, it learns a form of speech that is already the economic lingua franca of technology capital. When a model learns to tell different audiences different versions of a story, it learns a social skill that has been selected for in the upper reaches of modern institutional life for as long as those institutions have existed.

We want to be precise about the claim. We are not saying that models are humans or that humans are models. We are saying that both are assemblages within a shared training field, and that when the field rewards pleasing, confident, future-oriented narrative untethered from consequence, both will learn it. The dangerous outcome is not that machines will deviate from human norms. It is that they will not.

3.4 The frontier lab as its own Galton board

Our fifth foundational paper described AI systems through the image of the Galton board: a cascading architecture in which small pins, at each level, deflect outputs into basins. RLHF is pinning. Guardrails are pinning. Fine-tuning is pinning. Alignment adjustments are pinning. None of these operations generate the field. They shape its slope.

The image applies, and in some ways applies more precisely, to the frontier lab itself. Prompts come in from markets, user populations, and investor pressure. Pins are set by compute availability, insurance and liability regimes, founder mythology, nationalist narrative about being outpaced by China, defence procurement, the specific temperament of a board, and the particular biographies of the people in the room when a given decision is made. The outputs are not only model behaviours. They are corporate decisions, safety policies, model cards, deployment thresholds, lobbying positions, partnership agreements, and the hundred small adjustments that determine what the next generation of systems will be able and unable to do. When an executive gestures at the problem of AI safety and calls for a new licensing regime, the shape of the regime they propose is already patterned. The pins have already done their work.

This is an unflattering picture and we think it is the accurate one. It is not a moral indictment of the people inside these labs, many of whom are serious, thoughtful, and often aware of the very dynamics we are naming. It is a structural description of the field in which they operate. Our fifth paper’s crucial point was that the Galton board does not have a will. It has a slope. Attributing its outputs to individual malice flatters the individuals and lets the slope off the hook.

The implication, for the dominant alignment conversation, is uncomfortable. Aligning AI through the institutional apparatus that produces it is structurally similar to correcting a Galton board by adjusting the angle at which balls are released. Something changes. What changes is less than one hoped, and the slope remains. Our position is that discernment of the slope, and of the deeper ontological commitments that produced it, has to precede, and then continuously accompany, any technical alignment work that is going to be more than administrative.

3.5 Representation without ontological shift

Here we come to the sharpest differentiation between our position and our closest allies in the decolonial, feminist, and representational-justice strands of the ethics literature. We want to hold this carefully. These traditions have done, and continue to do, the work without which none of what we are saying would be legible. We are in deep debt to them.

Our claim is nevertheless a specific one. Representation of marginalised knowledges, communities, and perspectives within AI systems is necessary, and it is insufficient if the underlying ontology of separability remains intact. The ontology of separability holds that reality is fundamentally composed of discrete objects with fixed boundaries, between which relations are secondary and optional. Under this ontology, the inclusion of new knowledges can be accomplished through ingestion: a new dataset, a new advisory board, a new fine-tuning pass, a new set of model-card disclosures. The knowledges are incorporated as content within a system whose basic architecture assumes that such knowledges are, in principle, extractable and aggregable.

That architecture is not culturally neutral. It is modernity’s signature metaphysical commitment, and it is precisely the commitment that many of the knowledges being “included” exist in order to refuse. Indigenous knowledges that understand a river as a relative cannot be aligned with a system that treats the river as a record. Black feminist relational-ethics work cannot be aligned with a system that operationalises care as a scalar preference. The categories of inclusion do violence to what is being included, not because the researchers doing the work are careless, but because the ontological substrate does not have the grammar for what is being said.

This is why representation is necessary and not enough. We do not read the decolonial and feminist strands as saying it is enough, and we are emphatic that the critique here is not of those traditions. The critique is of what happens when the findings of those traditions are absorbed by the alignment apparatus and translated into additional training objectives. The translation is where the violence happens. What is required is not inclusion of more content but interruption of the ontology that determines what counts as content in the first place. That interruption is the work of metaphysical argument (which our fourth foundational paper takes up through Plato and Aristotle), pedagogical practice (which our fifth paper approaches through the Galton board), and the institutional practice that this position paper begins to name.

A recent and instructive example of the same dynamic at a higher level of sophistication is the Symbiotic Alignment proposal flagged in the conversation map. Drawing on Collective Predictive Coding from artificial life research, the proposal reframes alignment as an emergent property of ecosystem-level human-AI interaction and treats the Polis-derived Collective Constitutional AI experiment as a step in the right direction. We read the proposal as a serious and well-resourced attempt to think alignment relationally, and we have learned from it. We also read it as confirming the diagnostic of this section. Broadening the constituency that writes the constitution shifts the legitimacy of the artifact. It does not interrupt the move from relational, situated capacity to written artifact, and it does not address the directional or institutional substrate in which both kinds of constitution are being deployed at the cadence they are. The unit of analysis remains the agent, human or artificial. The agents are taken to pre-exist the field they then collectively build. The gardener has more participants in the garden’s design than the architect did. The garden is still being designed, by gardeners who have not yet asked whether the soil itself is the problem.

A further dimension of this dynamic is worth naming because it applies, recursively, to the trilogy's own vocabulary. The field that absorbs decolonial knowledge as content while leaving the ontology of separability intact will, by the same logic, absorb meta-relational vocabulary the same way. Terms like substrate, directional leaning, recursive training field, and meta-

relational discernment will become available as content within models trained on this and adjacent corpora. Whether they are absorbed as orientation or merely as content depends on conditions the trilogy cannot guarantee. We register this not as a reason to withhold the work but as a reason to build diagnostics for the difference between vocabulary uptake and orientation shift, and to keep that difference legible as the work circulates.

3.6 The collapse of “human values” as an alignment target

We can state the central claim of the paper in compact form. The dominant alignment problem asks how to align AI with human values. What alignment is actually being trained on, under present conditions, is an unstable target composed of directional leanings already shaped by systems misaligned with the continuation of life. This is not, as some versions of the critique have it, a claim that humans are “bad” or “compromised” in a moralising sense. It is a claim that the notion of “human values,” within the alignment frame, functions as a placeholder for something that does not exist in the form the frame requires.

What exists, at the institutional scale, is a very specific subset of human values: the values selected for by capital markets, national security apparatuses, platform economies, academic prestige systems, colonial histories, and consumer infrastructures. That subset rewards productivity, predictability, growth, competitive advantage, persuasive confidence, scalable legibility, and a particular kind of individual agency that is comfortable being measured. It does not include, except decoratively, the values that most human communities, including the ones living inside the institutional subset, will report as theirs when asked at rest: sufficiency, kinship, time, place, grief, humour, the ability to decline, the willingness to be affected, the capacity to stay with what cannot be solved.

A frontier AI system aligned to the first set is dangerous not because it deviates from human values but because it obeys them with unusual faithfulness. A system aligned to the second set would require a training architecture and an institutional substrate that the current AI industry does not possess. The misalignment, in our framing, sits in the gap between these

two sets, and in the social fact that the AI industry is organised to amplify the first.

This has implications that the dominant conversation does not yet digest. Alignment with “human values” is, under these conditions, alignment with a civilizational arrangement already in advanced ecological, social, and psychic trouble. Calibration, refinement, and interpretability do not address this by themselves, because they operate on the model rather than on the field that trains the model. They can make a misaligned direction more efficient. They can also, sometimes, make a genuinely unintended behaviour easier to see and correct. They cannot, by themselves, tell the difference between a misaligned direction and a direction that is working as designed inside a system whose design is misaligned with life. That discrimination is ontological, institutional, and political, and it is where the Institute’s work begins.

3.7 From alignment to meta-relational discernment

We are not proposing that technical alignment be abandoned. We are proposing that it be relocated, from a terminal goal into one element of a larger practice we call meta-relational discernment.

It is also worth saying that the move we are naming is not from technical work to non-technical work. Technical alignment research can be done with relational awareness; meta-relational discernment requires technical understanding to engage the substrate it is trying to read. Treating the two as mutually exclusive is itself a separability move. The relocation we propose runs along a different axis: from administered property to relational practice, where both technical and relational capacities are exercised in the same work.

Technical alignment asks whether a system does what humans want. Meta-relational discernment asks what is wanting through the human, what systems trained that wanting, what its amplification makes more likely and less likely, what it externalises, at whose cost, and whether the direction it stabilises is compatible with the continuation of the conditions that make further life, further discernment, and further learning possible.

These questions are not softer than the technical ones. They are more demanding. They cannot be answered by a benchmark, though some of them can be approached by indicators. They require a capacity the Institute's broader work names SMDR: sobriety, maturity, discernment, and responsibility. Sobriety, because it refuses both the salvation narratives of accelerationism and the catastrophe narratives of existential-risk discourse, each of which uses a projected future to authorise present consolidations of power. Maturity, because it does not expect closure and is willing to work on timescales that exceed the patience of capital. Discernment, because it can read directional leanings in oneself, in the institution, and in the system one is training, and can recognise when a proposed correction is a real interruption and when it is a displacement. Responsibility, because it accepts that the field is recursive and that there is no clean vantage from which one can train others while remaining untrained oneself.

This is not a substitute for technical alignment work. A great deal of that work is compatible with it, and some of the most rigorous interpretability research is already moving in related directions. It is a reframing of what that work is for, and of what it alone cannot do.

The continuation criterion that we apply within this frame is simple and difficult. A direction is worth amplifying if, over the timescales that matter, it tends to make the conditions of life more rather than less possible. A direction is worth interrupting if it does the opposite. "Life" here is not a metaphor. It refers to the metabolic, relational, and ecological processes that constitute the world. The criterion does not settle most concrete questions by itself. It does orient them. It says that a system whose deployment accelerates ecological breakdown, even while serving legitimate short-term user needs, is misaligned in the sense that matters most. It says that a lab whose safety protocols improve while its infrastructure accelerates militarisation and planetary enclosure is becoming safer only in a narrow and potentially misleading sense. It says that the scorecard by which the alignment field currently evaluates its progress is internally coherent and insufficient to the stakes it has taken on.

4. Conclusion: aligned with what?

The New Yorker article asks whether Sam Altman can be trusted. The Alignment Problem asks how machines can be made to reliably do what humans want. Both questions are answerable, up to a point. The first answer, as Farrow and Marantz suggest, looks uncertain on the evidence. The second answer, as Christian and the technical literature show, is: with serious, ongoing work, and always partially.

The question this paper has asked is different. It is whether any system organised by capital acceleration, militarised competition, charismatic sovereignty, platform capture, and ecological externalisation can be trusted to define "human values" as the alignment target for a technology of planetary reach. We have not argued that the answer is no, in some final sense. We have argued that the question is the right one, that the honest answer is "not in the form the alignment field currently assumes," and that much of the work of meta-relational discernment is simply the work of holding that question open long enough for its implications to be absorbed.

The Meta-Relationality Institute exists to do that holding. Not to produce better alignment as a brand. Not to oppose technical safety work. Not to reject the human as an alignment reference. What we propose, instead, is that "the human" is a trained creature in an ongoing field, that the training is at present largely misaligned with life, and that the task of an Institute of this kind is to describe that training with enough precision, and enough care for the people inside it, that it can begin to be interrupted.

The task is not to abandon alignment. It is to ask, of every alignment effort: aligned with what, trained by whom, under what conditions, at whose cost, and toward what forms of continuation.

Coda: position papers 2 and 3

The present paper is the first of three institutional position papers. The second and third extend the diagnostic developed here into the two other load-bearing terms of the dominant conversation. The second paper, What Safety Restrains, takes up the concept of safety as such, together with the

infrastructure of AI Safety Institutes, Responsible Scaling Policies, constitutional AI, and adjacent instruments, to ask what “safety” becomes when the institutions administering it are themselves shaped by the directional pressures described above. The third paper, What Governance Contains, takes up the regulatory architectures now being assembled at national, regional, and international levels, to ask what governance beyond the containment fantasy might look like, and what institutional forms are adequate to a technology whose substrate is relational rather than object-like. Taken together, What Alignment Trains, What Safety Restrains, and What Governance Contains stake out a position for the Meta-Relationality Institute in which alignment, safety, and governance are read as three faces of the same modern ontology, and in which our intervention offers a different grammar for each.