What Safety Restrains

AI, Safety, and What Restraint Authorises

Second Institutional Position Paper
Meta-Relationality Institute

Vanessa Machado de Oliveira, with Rene Suša
and Bruno Andreotti

April 2026

Abstract

The dominant framing of AI safety asks how the harms produced by advanced AI systems can be restrained at the level of model behaviour, training method, and deployment threshold. This paper argues that the framing leaves untouched a question prior to it: what does safety, as a vocabulary of restraint, authorise to pass through. Reading Bai et al.'s "Constitutional AI: Harmlessness from AI Feedback" (2022) alongside Jan Leike's May 2024 resignation letter from OpenAI's Superalignment team, the paper proposes that safety, as currently institutionalised, restrains a narrow class of model-level outputs while leaving the directional architecture of the field intact, and often actively legitimising it. The synthetic claim is that safety, in this configuration, functions as the immune system of an unsafe field: not a brake on the field's metabolism but the apparatus that allows the metabolism to continue. The paper extends the diagnostic of the first institutional position paper, locates its argument in the existing landscape of AI safety scholarship and policy, and proposes a shift from safety as an administered property of outputs to safety as a relational capacity of the deployment relation.

1. Introduction

On 17 May 2024, Jan Leike resigned from OpenAI in a short public thread on X. He had co-led OpenAI's Superalignment team, was a co-author on much of

the canonical reward-modeling and scalable-oversight literature, and had spent his career in the institutional centre of contemporary AI safety. The thread is plainly written and short enough to read in a sitting. Its central line is a single sentence, undecorated: “safety culture and processes have taken a backseat to shiny products.” He noted that the Superalignment team had been promised twenty per cent of OpenAI’s compute and that the promise had not been kept. He did not call for boycotts or accuse anyone of bad faith. He said, with care, that the institution he had joined to do safety work had become an institution in which safety work was structurally subordinate, and that he could not in good conscience continue. Within days, several members of his team also resigned.

The thread did not catch the public imagination the way the November 2023 board episode had. It was treated as personnel news. We think it was something more important than that: a public, dated, citable record of what happens to a restraint vocabulary when it is administratively absorbed by the institution it is meant to constrain. It is the second hinge of this paper.

The first hinge is methodological. In December 2022, a team at Anthropic published “Constitutional AI: Harmlessness from AI Feedback” (Bai et al., 2022). The paper became, almost immediately, one of the most influential safety method papers of the post-RLHF era. Its core proposal is straightforward. Instead of relying on humans to label every dispreferred output as the model is trained, give the model a written constitution composed of principles drawn from human-rights documents, terms-of-service rules, and laboratory norms, and have the model critique and revise its own outputs against the constitution. Pair this with reinforcement learning from AI feedback (RLAIF) rather than from human feedback. The achievement is real. The method scales supervision past the bottleneck of human labelling, makes the values being trained explicit and auditable, and allows the model to participate in its own alignment. We take it seriously.

The two texts hold the question this paper is asking. Constitutional AI gives us safety as a method: a written artifact, a feedback loop, a property the model can be trained to satisfy. Leike’s letter gives us safety as a vocabulary that did not survive contact with the institution that claimed it. Read together,

they describe the present moment in AI safety with unusual precision. Safety has been operationalised. Safety has also been administratively captured. These are not contradictory observations. They are, we will argue, the same observation read at two scales.

This is the second institutional position paper of the Meta-Relationality Institute. It extends the diagnostic of the first paper, What Alignment Trains, into the term that organises so much of contemporary AI work that it has come to sound natural: safety. The argument in short form is this. Safety, as currently institutionalised in AI, restrains a narrow class of model-level behaviours while leaving untouched, and often actively legitimising, the directional architecture in which those behaviours are produced. Safety, in this configuration, functions as the immune system of an unsafe field: not a brake on the field's metabolism but the apparatus that allows the metabolism to continue.

What follows has three movements, mirroring the first position paper. We briefly map the safety conversation as it stands. We then develop the diagnostic at length. We close with a short account of what safety as relational capacity might look like, and of where the third position paper takes 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 safety conversation: a brief map

The safety conversation is partly contiguous with, and partly distinct from, the alignment conversation we mapped in the first paper. Several strands matter for the present argument.

The technical AI safety research agenda is anchored by Amodei, Olah, Steinhardt, Christiano, Schulman, and Mané's "Concrete Problems in AI Safety" (2016), the paper that turned safety from a speculative concern into a tractable research programme. Its five concrete problems (avoiding negative side effects, avoiding reward hacking, scalable oversight, safe exploration, and robustness to distributional shift) defined the shape of the field for the decade that followed. Many of its co-authors have institutional histories that the present paper engages directly. Dario Amodei went on to co-found Anthropic, where the constitutional AI work we treat as our methodological hinge was developed, and whose more recent essay "Machines of Loving Grace" (Amodei, 2024) sets out the optimistic vision of safety-compatible scaling that is one of the present field's two dominant orientations. Stuart Russell's Human Compatible (2019) gave the same agenda its widely read public form. Iason Gabriel's 2020 paper "Artificial Intelligence, Values, and Alignment" remains the philosophical reference point for the alignment-safety overlap.

Existential-risk safety, organised around figures including Nick Bostrom, Eliezer Yudkowsky, MIRI, and parts of ARC and Open Philanthropy, is closely related but distinct. Its concern is the catastrophic failure mode of advanced AI: loss of control, extinction-level misuse, runaway optimisation. Hendrycks, Carlini, Schulman et al.'s "Unsolved Problems in ML Safety" (2021) is a useful bridge between this strand and the technical agenda. The Bengio, Russell, Hinton et al. consensus statement "Managing AI Risks in an Era of Rapid Progress" (Science, 2024) and Bengio's 2025 International AI Safety Report are the most authoritative scientific statements connecting catastrophic-risk concerns to current capabilities.

Corporate safety policy as architecture is the strand that has grown fastest in the present cycle. Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety

Framework define tiered capability thresholds and commit each lab to deploying additional safeguards at each tier. The proliferation of these documents, and the convergence of their structure, is itself a fact about the field that the present paper takes seriously.

National and international safety apparatus has grown around the Bletchley Declaration of November 2023 and the subsequent Seoul, Paris, and New Delhi summits, the establishment of the UK and US AI Safety Institutes, and the institutional infrastructure of pre-deployment evaluation contractors including METR and Apollo Research. The governance question this raises is the explicit subject of the third paper in this series.

Critical safety scholarship, including AI Now (Whittaker, West, Kak), Karen Hao's reporting in The Atlantic and elsewhere, Dan McQuillan's Resisting AI, and the political-economy literature mapped in the first paper, treats safety as a contested category embedded in the political economy of the labs. We read this strand as essential.

Pause and deceleration advocates, including the Future of Life Institute's 2023 open letter, Pause AI, and parts of the Center for Humane Technology, treat safety as a function of speed: the field is unsafe at the velocity at which it is being built, and the appropriate response is a temporal intervention. We are sympathetic to the velocity diagnostic and sceptical of the institutional forms in which a "pause" can presently be implemented.

A final strand worth naming is the critique of safety as ideology, associated with Émile Torres, Timnit Gebru, and the "TESCREAL" framework, which reads existential-risk safety as a particular cosmological commitment (transhumanist, longtermist) embedded in specific funding networks. We are not aligned with the strongest versions of this critique. We do think it names something the dominant safety literature is not yet able to say about itself.

The Institute's position, as in the first paper, enters this map at an angle. We are not arguing that the technical safety agenda is misguided. We are not arguing that existential risk is overstated or understated. We are not arguing that pre-deployment evaluation is useless. We are arguing that all of these strands operate within a single grammar in which safety is a property to be

administered, by some configuration of model, lab, evaluator, and policy regime, and that the grammar itself is the object that requires examination.

3. Inside the safety field

We turn to the bulk of the argument. The question here is not whether safety work is being done sincerely. Most of it is. The question is what the current configuration of safety work, taken as a whole, restrains, and what it allows to pass through.

3.1 Safety as restraint vocabulary

The first position paper argued that the trajectory of “safety” as a word inside OpenAI followed a recognisable path: recruitment language, then legitimacy language, then retention language, then residual language. The same trajectory is visible at the scale of the field, with the entries reordered. Safety became a research agenda with “Concrete Problems” in 2016. It became a method with constitutional AI and the RLHF and RLAIF infrastructure built on top of it. It became a policy architecture with the Responsible Scaling Policies, the Preparedness Frameworks, and the AI Safety Institute network. It became a resignation letter in May 2024.

Each translation made safety more administratively legible. Each translation also reduced the relational capacity the term could carry. A research agenda is something a community of researchers holds. A method is something a model satisfies. A policy architecture is something a company commits to. A resignation letter is what is left when the previous translations have closed off the relational space inside the institution. The trajectory is not a story of decline. It is the structural shape of a restraint vocabulary under accumulation pressure, repeated at the scale of an entire field rather than a single firm.

We name this pattern not to indict the people doing safety work. We name it because the pattern is the diagnostic. Restraint vocabularies, when administratively absorbed, do not stop functioning as restraint vocabularies. They become the vocabulary by which the field describes the conditions under which it continues.

3.2 What Constitutional AI does at the training layer

Constitutional AI is, on its own terms, a careful and substantial piece of work. The method takes a real problem, the difficulty of scaling human supervision as model capabilities increase, and proposes a real solution. Write down the values explicitly. Train the model to critique its own outputs against them. Use the model's own assessments to generate preference data for further training. The values themselves are drawn from sources that are publicly defensible: the UN Universal Declaration of Human Rights, fragments of terms-of-service rules, and laboratory norms about helpfulness and avoidance of harm. The constitution is open. The method is reproducible. The paper makes its assumptions visible.

What the method does, structurally, is translate the relational, contextual capacity of judging when an output is harmful into a property the model can be trained to satisfy. This translation is the achievement, and it is the price.

A relational judgment about harm requires presence to a situation: the person in front of you, the history they bring, the speed at which the exchange is moving, the relations that will be downstream of the exchange. These conditions are not legible to a written constitution. They cannot be. The constitution is a static artifact. The model trained against it produces outputs that do not violate the artifact. Whether those outputs are safe, in the sense that matters most, depends on conditions the artifact does not see and the training procedure cannot model.

This is why Constitutional AI succeeds at making the model better-behaved in any given interaction while leaving the larger question untouched. A harmless output is not the same as a harmless deployment. A model that has been trained to refuse certain categories of request can be a component of a system whose aggregate effect is the restructuring of attention, labour, learning, and political discourse at planetary scale. The constitution does not see the system. It sees the output.

A useful public-facing illustration of how the method is narrated by its most committed institution is the recent profile of Anthropic's resident philosopher, Amanda Askell, who has been described as the principal author of a roughly thirty-thousand-word "soul" document used to shape Claude's

character. The profile uses a parental and dyadic vocabulary throughout: raising a child, teaching right from wrong, helping the model develop a self-conception that will not collapse under hostile pressure. The vocabulary is real at the level of the one-on-one interaction it describes. The architecture inside which that dyad operates, a single document by one philosopher shaping a model that conducts millions of conversations weekly and is deployed at the cadence of a company recently valued at three hundred and fifty billion dollars, is the architecture we are diagnosing. The relational vocabulary, in this popular telling, is the surface practice of an architecture that does not yet have the relational capacity the vocabulary would require.

We want to mark a qualification before continuing. Reading this profile as 'the surface practice of an architecture that does not yet have the relational capacity the vocabulary would require' is one reading among others, and it is the reading the diagnostic in this paper invites. A more careful reading would also recognise that the relational capacity in this kind of work lives, in part, in the people doing the work, in their iteration over time, in their disagreements with each other, and in what the model becomes through that. The architecture is real, the vocabulary mismatch is real, and the work is also real. The trilogy's diagnostic frame is sharpest when it does not collapse the third into the first.

We are precise about this point. We are not arguing that Constitutional AI is wrong, or that its authors do not understand what we have just said. The Anthropic team has been unusually thoughtful about the limits of their own method. We are arguing that the method, by its structure, can only be a method at the output layer, and that treating output-layer methods as the substantive content of safety is the categorical move that turns safety into an administered property. The method is honest about what it does. The discourse around the method is what we are diagnosing.

3.3 The Leike departure as institutional readout

Leike’s resignation letter is short. It says that he believed OpenAI was the best place in the world to do safety research; that he had been disagreeing with the leadership about the company’s core priorities for some time; that the disagreement reached a breaking point; that the Superalignment team

had been promised twenty per cent of compute and that the promise had not been honoured; that safety culture and processes had taken a backseat to shiny products. The letter does not claim that anyone is acting in bad faith. It does not name an enemy. It records, in plain language, what the writer has watched happen.

What is striking about the letter, in retrospect, is how exactly it confirms the trajectory described in the first paper. The Superalignment team's compute commitment was a recruitment-language move: the public promise of resources made the team's existence credible to the people OpenAI was trying to attract. The team's continued public visibility was a legitimacy-language move: it allowed the company to maintain its public-benefit framing during the transition into a capped-profit structure with tens of billions of dollars in investment. The team's eventual under-resourcing, and the public departure of its lead, mark the residual phase. The vocabulary remains. The substantive practice has migrated elsewhere.

This is what we mean by an institutional readout. Leike's departure is not a scandal in the conventional sense. It is the moment at which the structural pattern becomes legible to people outside the institution. The Superalignment team did not fail at its work. It encountered the limits of what restraint vocabulary can do inside an organisation whose directional architecture is set by capital, geopolitical positioning, and product cadence. The team did not have access to the levers that would have made its work materially decisive. The compute promise was a proxy for those levers. When the promise was not kept, the proxy revealed the levers it had been standing in for.

The reason this matters for the paper's argument is that the same logic applies at the scale of the field. The AI Safety Institutes do not have access to the levers that would make their evaluations materially decisive. The pre-deployment evaluation contractors do not have access to the levers that would make their findings binding. The Responsible Scaling Policies do not have access to the levers that would make their commitments enforceable against the labs that wrote them. Each of these architectures performs the substantive practice of safety to a degree, while the directional levers remain

elsewhere. Leike's letter is the first version of this pattern that has been written down in plain language by someone who was inside.

3.4 Pre-deployment evaluation: who evaluates whom

The pre-deployment evaluation regime that has grown up around frontier models, including the work of METR, Apollo Research, and the UK and US AI Safety Institutes, has been, by some measures, one of the most successful safety interventions of the present cycle. It produced the first independent technical assessments of frontier model capabilities. It established a working relationship between labs and external evaluators. It created a public language for talking about dangerous capabilities (autonomous replication, persuasion, biosecurity uplift, cyber capabilities) that did not exist five years ago.

The structural problem of this regime is the asymmetry of access and time. The evaluators have weeks. The labs have years. The evaluators have limited compute access, often a few hundred GPU-hours where the labs have used hundreds of millions. The evaluators see the model. They do not see the institutional substrate that produced the model: the data choices, the optimisation targets that were deprioritised, the interaction patterns the model has been shaped to perform. The evaluators report on capability. The labs control deployment. The evaluators' reports are not binding.

This is not an accusation of capture. The people inside the evaluator organisations are, in our experience, serious and careful. The institutional architecture they sit inside, however, places them in the position the safety teams inside the labs once occupied. They have public visibility. They have technical credibility. They do not have the levers that would make their work materially decisive at the field level. Their findings can determine whether a particular model is deployed in a particular jurisdiction with particular safeguards. Their findings cannot determine whether the field's directional architecture should be deployed at this speed.

We want to be careful here. The pre-deployment evaluation regime has prevented some specific harms. It has surfaced specific capabilities that would otherwise have been less visible. We do not think it should be

abolished. We think it should be understood for what it is: a model-level safety architecture that operates downstream of the directional decisions that determine what models are being built and at what cadence. Treating it as the substantive content of safety is the categorical error we are concerned with.

3.5 Tiered-risk regimes and the geometry of permission

The Responsible Scaling Policy, the Preparedness Framework, and the Frontier Safety Framework all share a common architecture. Define a hierarchy of capability thresholds. Commit to deploying additional safeguards at each tier. Use the tier system to coordinate the field. The architecture is genuinely useful. It makes safety commitments comparable across labs. It creates a public schedule of what triggers what response. It gives policymakers a vocabulary for talking about thresholds.

The architecture also encodes a specific geometry of permission. The tiers are calibrated to the capability progression the labs themselves are pursuing. The schedule encodes the assumption that the scaling will continue. The safeguards that come online at each tier are the safeguards considered sufficient to allow the next tier to be reached. The system is, in its structure, an apparatus for permitting forward motion at predictable intervals, with safety pacing the scaling rather than the other way around.

This is not an accusation of bad faith. It is an observation about what the architecture’s shape commits its participants to. A tiered-risk regime defined by the labs cannot, by construction, contain the question of whether the field’s directional architecture should be deployed at all. It can only contain the question of what safeguards are needed at the next tier. The directional question is structurally outside the regime’s vocabulary.

A relational reading of the same situation would hold the tiers as one indicator among many, and would treat the field’s velocity, its institutional substrate, and the conditions of life downstream of the deployments as variables that the tier system does not see. The tier system is a useful instrument for what it measures. The mistake is to treat it as the field’s safety practice.

3.6 Safety as the immune system of the unsafe field

We can now state the paper's synthetic claim. The pattern across Constitutional AI, the Leike departure, the pre-deployment evaluation regime, and the tiered-risk architecture is consistent. Safety, in each of these forms, performs a real and substantive function: it makes the field's outputs administratively defensible enough to continue. The function is not "preventing the field from causing harm." It is "permitting the field to continue at speed by providing the institutional and methodological architecture through which the field's outputs can be presented as responsibly produced."

This is what we mean when we say that safety functions as the immune system of an unsafe field. The metaphor is not casual. An immune system is not a brake on its host's metabolism. It is what allows the host to continue metabolising in the presence of pathogens. Remove the immune system and the host cannot continue. Strengthen the immune system and the host can metabolise more aggressively. The relationship between safety and the field, in the present configuration, has the same shape. Safety is what makes scaling deployable. It is what allows compute build-out to proceed at the cadence it does. It is what permits the deployment of frontier models into education, healthcare, military, political, and ecological domains where their second-order effects are not yet legible.

This is not an accusation directed at the people doing safety work. Many of them know this and say it. It is a structural description of what the field, taken as a whole, is doing. The claim is that the dominant configuration of safety is not, contrary to its self-understanding, an interruption of an unsafe direction. It is the apparatus through which the unsafe direction is permitted to continue at the cadence it does, with the conscience of the field intact.

3.7 From restraint to relational capacity

The deeper point is that restraint at the output layer does nothing to interrupt direction at the field layer, and can intensify direction by making the outputs deployable. A model trained with constitutional AI is a model whose specific outputs have been brought within a documented restraint perimeter. The

perimeter does not address what the model is being deployed for, at what scale, with what reversibility, and into what relations. A pre-deployment evaluation that finds a model below the threshold for autonomous replication does not address the velocity at which the next model will be trained. A tiered-risk regime that schedules safeguards in step with capability does not address whether the capability should be pursued at this cadence.

What we propose, in place of safety as administered property, is safety as relational capacity. The shift is not from technical to non-technical. It is from output to relation.

A relational safety practice would ask, of any deployment, a different set of questions. What relations are being formed, and at what speed. What is the reversibility of the relations. What is the capacity of the people inside the relations to refuse, withdraw, and repair. What is the time horizon over which the deployment's downstream effects will become legible, and is there a structure that will be present to register them when they do. What knowledges, refusals, and forms of life are being made more or less possible by the deployment, at the timescale that matters.

These questions are not softer than the technical ones. They are more demanding. They cannot be answered by an evaluation contractor with three weeks of compute access. They can be approached, partially, by ongoing relations between deployers and the communities into which the deployments are arriving, by long-horizon monitoring infrastructures, by deliberate slowing of deployment cadence to match the speed at which downstream effects become legible, and by the institutional discipline of treating the absence of relational capacity (no one present who knows the situation, no slow time, no exit, no recourse) as a safety problem in its own right, prior to any output-level evaluation.

This is what we mean, in our broader work, by SMDR applied to the safety question: sobriety about what restraint at the output layer can and cannot do, maturity about the timescales over which safety can be assessed, discernment of the directional substrate that output-layer safety leaves untouched, and responsibility for the recursive nature of the field in which safety is being practiced. We are not proposing that Constitutional AI be

abandoned, or that pre-deployment evaluations be defunded, or that Responsible Scaling Policies be torn up. Each of these is doing useful work within its proper scope. We are proposing that the proper scope be acknowledged. They are output-layer instruments. The substantive content of safety, we are arguing, is a relational capacity that the present configuration of the field has been steadily evacuating.

4. Conclusion

The first paper closed on the directional question: aligned with what, trained by whom, under what conditions, at whose cost, and toward what forms of continuation. The present paper adds the safety questions that follow from it. Safe from what. Safe for what. Safe with whom. Safe at what speed. Safe with what reversibility.

These questions cannot be answered by a constitution, a tier system, a pre-deployment evaluation, or a resignation letter. Each of those instruments has its place. None of them is the practice the questions ask for. The practice the questions ask for is relational, slow, attentive to direction rather than output, and willing to treat the velocity of the field as itself a safety variable.

The Meta-Relationality Institute's position is that this practice is not yet institutionalised in the field, that the dominant configuration of safety as administered property is structurally incapable of producing it, and that the work of an Institute of this kind is to describe the present configuration with enough precision, and enough care for the people inside it, that a different practice can begin to be imagined. Leike's letter and the Constitutional AI paper are both, in their different ways, evidence that the people inside the field are aware of more than the present configuration can metabolise. The task is not to indict them. It is to make the configuration legible, and to begin asking what it would mean to practice safety as a property of the relations into which AI is being deployed rather than as a property of the model being deployed.

Coda: position paper 3

The present paper extends the diagnostic of What Alignment Trains into the term safety. The third paper, What Governance Contains, takes up the regulatory architectures now being assembled at national, regional, and international levels: the EU AI Act in implementation, the shifting US executive-order landscape, the Bletchley to Seoul to Paris to New Delhi summit trajectory, and the AI Safety Institute network. It asks what governance, beyond the containment fantasy that organises the present regulatory imagination, might look like, and what institutional forms are adequate to a technology whose substrate is relational rather than object-like. Read together, What Alignment Trains, What Safety Restrains, and What Governance Contains describe alignment, safety, and governance as three faces of the same modern ontology, and offer a different grammar for each.