Some[Body] Must Receive That Pain for Agent Accountability

Botao Amber Hu
University of Oxford
Oxford, UK
botao.hu@cs.ox.ac.uk

Helena Rong
New York University Shanghai
Shanghai, China
hr2703@nyu.edu

Abstract

AI agents increasingly act consequentially in the real world. This creates a problem we call *consequence reception*: harm occurs, the producing system is identified, yet no continuing agent receives consequences in a way that changes future behavior. Pain, understood mechanistically as a corrective feedback signal, is foundational to canonical theories of punishment—deterrence, rehabilitation, retribution, and incapacitation all assume a continuing locus that registers the signal and updates behavior. That, in turn, requires a body for the signal to land on: a boundary whose integrity it protects, a locus where it accumulates, consolidation that converts episodic signal into durable update, and a substrate that responds by altering future action. Current LLM agents—software-defined composites of weights, prompts, tools, memory, and credentials, freely swapped, copied, reset, and reassembled—satisfy none of these conditions. The two prevailing legal responses therefore fail to achieve consequence reception. The thin-identity agent-principal dyad has a body but no *consequence–agency coupling*: the human bears pain for behaviors beyond their control—Elish's *moral crumple zone*. The thick-identity Arbel et al.'s *Algorithmic Corporation* creates legally legible entities but does not guarantee that any AI decision architecture receives pain as a behavioral signal. Achieving consequence–agency coupling is therefore a sociotechnical infrastructural problem, not only a legal one. Until such architectures exist, high-stakes AI deployment should remain tethered to accountable human principals with meaningful control, proportional liability, and authority to constrain or terminate the agent. *If some body does not receive the pain by design, some body will receive it by default.*

1 Introduction

On March 18, 2018, an Uber autonomous test vehicle struck and killed Elaine Herzberg as she crossed a road in Tempe, Arizona. The vehicle's perception stack had detected her 5.6 seconds before impact and successively classified her as an unknown object, a vehicle, and a bicycle; the emergency-braking subsystem had been disabled to prevent erratic behavior during testing [National Transportation Safety Board, 2019]. The safety driver, Rafaela Vasquez, was charged with negligent homicide; in 2023 she pleaded guilty to endangerment and was sentenced to three years of supervised probation. Uber faced no criminal liability. The decision substrate—the perception stack, classification algorithm, and control software—received nothing. The human positioned to monitor it received the criminal record.

This is what Elish [2019] calls a *moral crumple zone*—a human positioned in a technical system to absorb moral and legal responsibility for failures whose proximate causes lie in components they cannot meaningfully control. It is also, with refinement, the structure proposed for AI-agent accountability by the thin response: pair each AI agent with an identifiable human or organizational

Preprint.

principal and treat the pair as the unit of accountability [Shavit et al., 2023, Chaffer, 2025], as in the EU AI Act's provider/deployer split. The principal is meant to supply the body the agent lacks.

We call this gap *consequence reception*: the property whereby a sanction imposed on a system produces a durable change in the substrate that generates its future behavior. Reception is distinct from *attribution*, which identifies the producing system, and from *feedback*, which is any signal returned to it. The gap itself is not new—Matthias [2004] identified the "responsibility gap" created by learning automata, Danaher [2016] showed it is especially acute for retribution, and Santoni de Sio and Mecacci [2021] taxonomize four distinct responsibility gaps. Our contribution is to identify what *substrate property* must be present for any of these gaps to be closed. Most current AI governance discourse operates at the attribution level. Some, including post-deployment fine-tuning and reinforcement learning from human feedback [Christiano et al., 2017], operates at the feedback level. Almost none operates at the level of reception, because reception requires properties of the substrate that contemporary AI agents do not possess.

This paper makes three claims. First, accountability requires consequence reception: the canonical theories of punishment—deterrence, rehabilitation, retribution, and incapacitation—each presuppose a continuing locus that registers consequences as a behavioral signal and updates accordingly. Second, reception requires a substrate we call, functionally, a *body*: an architecture satisfying four conditions—boundary, locus of accumulation, consolidation, and substrate response—which together exhaust what the punishment theories need to operate. Third, both prevailing legal responses to AI accountability fail to achieve consequence–agency coupling. The thin agent–principal dyad supplies a body without coupling: humans absorb sanctions for behaviors they cannot meaningfully control, producing what Elish [2019] calls a *moral crumple zone*. The thick Algorithmic Corporation [Arbel et al., 2026] supplies legibility without coupling: a legal envelope wrapped around an architecture that does not receive. Closing the loop is therefore a sociotechnical problem, not a purely legal one.

The contributions follow this argument. We reframe pain mechanistically as a corrective feedback signal, sufficient to operate the canonical theories of punishment without taking a position on phenomenology. We develop a four-condition diagnostic that locates where existing accountability proposals succeed and where they fail. We argue that recent interpretability findings on functional emotion concepts in production language models [Sofroniew et al., 2026] open a research path toward consequence-receiving architectures, while sharpening the suffering-risking concern that such architectures raise. And we propose a conservative near-term deployment rule: until coupling-capable architectures exist, high-stakes deployment must remain tethered to human principals with three properties that current proposals do not require—meaningful control, proportional liability, and authority to terminate.

The framework operates at the mechanistic level and is deliberately silent on whether consequence-receiving systems would be conscious, would have moral status, or would suffer. These questions are entangled with our framework rather than orthogonal to it—what counts as a boundary depends partly on what is inside it—but governance cannot wait for their resolution. Section 6 argues that the framework is compatible with multiple resolutions of the underlying philosophical questions, including positions that recommend non-deployment.

2 Consequence Reception

2.1 Accountability, Skin in the Game, and Consequence Reception

Accountability is not punishment, trust, nor alignment. Following Bovens [2007], we understand it as a structured relationship in which an identifiable actor must justify its conduct to a forum, and the forum holds authority to impose consequences that durably constrain the actor's future behavior. Two features do most of the work. The first is answerability: the actor must be capable of being identified, examined, and held to account. The second, more demanding, is enforceability: the forum's consequences must in fact constrain. Answerability without enforcement collapses into oversight without teeth [Schedler, 1999, Grant and Keohane, 2005]. What distinguishes accountability from mere transparency is the credible expectation that consequences will land and continue to matter.

Enforceability presupposes something about the actor on whom consequences will land. Locke [1694] grounded personal identity as a forensic notion: identity exists, in the philosophical sense, precisely to underwrite the claim that the person who acted then is the person who answers now. Parfit [1984]

sharpened the formal requirement: identity is by definition a one-to-one relation, so any process that makes an actor copyable, branchable, or substitutable fractures the conditions under which sanctions can target the right entity. The institutional dual is well-documented. Friedman and Resnick [2001] formalized the cheap-pseudonyms problem: when identifiers can be cheaply abandoned and replaced, reputation systems cease to function as accountability mechanisms. Douceur [2002] established the corresponding cryptographic result: the cost of generating new identities determines whether identification can support security at all. Taleb [2018] states the unifying insight in institutional-economic terms: accountability requires *skin in the game*, which presupposes an identity that is hard to copy, hard to reset, and expensive to abandon.

We adopt the term *non-fungible identity* for the property the literature has been triangulating: an identity that satisfies all three of these conditions. Non-fungible identity is the institutional precondition of enforceability. Without it, sanctions can be evaded by substitution; reputational signals fail to bind future behavior; the structural relationship Bovens describes loses its grip on the actor it is meant to govern. Here is the paper's first analytic move. Non-fungible identity is necessary for accountability, but it is not sufficient. An actor can be perfectly non-fungible—uniquely identified, registered in a public ledger, bound to a cryptographic key, subject to legal recognition—and still fail to be accountable, because the substrate that generated the action does not update in response to the sanction. The institutional condition addresses *who* the actor is. It does not address whether the actor can *receive* anything as a behavioral signal. A non-fungible label attached to a system whose internal state cannot consolidate sanction is a label, not an accountability mechanism.

The missing condition is consequence reception: feedback that lands on a non-fungibly-identified actor, accumulates over time, consolidates into durable structural update, and alters the substrate that produces future action. Reception is not merely online learning from negative reward. A gradient update to a loss function satisfies substrate response in isolation, but without non-fungible identity the update can be evaded by forking, and without consolidation it can be overwritten by the next batch. Reception requires that the *same continuing entity* that acted is the one whose future behavior is durably altered—a conjunction of identity persistence and substrate update that no current deployment architecture guarantees. Reception and non-fungible identity together—not either alone—are what accountability requires. This refinement preserves the accountability literature's diagnosis while completing it. Scholars from Locke through Taleb have correctly identified the institutional condition; in the AI-agent case, the substrate condition becomes independent and demands separate treatment. The remainder of this section develops what reception requires of a substrate.

2.2 Punishment theories presuppose reception

The case for reception is implicit in every canonical theory of why punishment is justified. We make the implicitness explicit. The four theories that have organized two centuries of penal philosophy each presuppose, at the level of mechanism, that sanctions land on a continuing system that registers them and updates.

Deterrence, as Becker [1968] formalized it, is forward-looking and probabilistic: the prospect of sanctioned outcomes alters expected-utility calculations and shifts behavior at the margin. For this mechanism to operate, the future self contemplating action must anticipate consequences that will fall on it—which presupposes that the present self can in some manner store the prospect of those consequences. Without accumulation, no future self exists who could be the one anticipating; without consolidation, no anticipation persists across the gap from one decision to the next. Rehabilitation, in the paternalist framing of Morris [1981], is more directly mechanistic. Its goal is dispositional modification of the offender. A theory of rehabilitation that does not presuppose substrate response is incoherent: the modification has nowhere to land. Retribution, in the foundational treatment of Hart [1968], requires that the agent who acted is the agent who bears. The expressive variant due to Feinberg [1965] adds that punishment communicates moral judgment, which presupposes a recipient capable of being addressed. Either way, retribution demands boundary and locus: the very thing that makes an action attributable is the thing that must persist to bear the response. Incapacitation [Zimring and Hawkins, 1995], the most forward-looking and least morally loaded of the four, requires that the entity whose future capacity is constrained is the same entity that would otherwise act. Without identity continuity across the constraint, what is incapacitated is not what would have offended.

These are not four independent theories that happen to share a presupposition. Reception is the substrate beneath all of them: deterrence acts through it, rehabilitation operates on it, retribution

requires it, incapacitation modifies what it has shaped. Taleb and Sandis [2014] formulate the same requirement in institutional-economic terms: effective accountability requires symmetry between decision authority and exposure to downside, enforced through absorbing states that cannot be evaded. Whether expressed in penal-philosophical or in institutional vocabulary, the demand is identical. The theoretical apparatus of accountability rests on a substrate condition that has not, to our knowledge, been examined directly in the AI-agent context.

2.3 Pain is the feedback signal that makes punishment work mechanistically

What, mechanically, is the signal through which sanctions become substrate change? In humans, the signal we know best is pain. We retain the term but redefine it operationally. Pain in our usage is not phenomenal suffering and not merely a loss term in an objective function. It is the mechanism by which a continuing system encodes a consequence so that future action is altered. The question Bentham asked—*can they suffer?*—is the wrong question for our purposes. The question is whether they can receive, where receiving is the operationally specified joint occurrence of (i) a signal landing on a bounded continuing entity, (ii) accumulating over time, (iii) consolidating into durable structural change, and (iv) altering the substrate that produces future action. We bracket the question of phenomenal experience. Whether mechanistic reception requires phenomenology, and whether building reception-capable systems creates moral patients, are open questions we return to in Section 6.

Three lines of evidence converge on this operational definition. The first is reinforcement signaling. Sutton and Barto [2018] formalize how reward and punishment shape policy through weight updates, and the biological substrate of this learning has been mapped to dopaminergic prediction-error signals [Schultz et al., 1997]. Seymour [2019] synthesizes these threads into a unified computational account: pain recasts from intangible subjective experience to a precise, objectifiable control signal for reinforcement learning—precision-weighted, prediction-error-based, and functionally necessary for adaptive avoidance. The behavioral signature of reception is asymmetric: organisms update more strongly from received losses than from equivalent gains [Kahneman and Tversky, 1979, Tom et al., 2007], and learning from aversive outcomes depends on durable memory consolidation mediated by stress hormones [McGaugh, 2015]. Jepma et al. [2018] show the asymmetry at the neural level: expectancy effects on pain are self-reinforcing, with aversive prediction errors producing stronger behavioral updating than disconfirming ones. Pain functions as a privileged feedback signal in part because it is hard to ignore: it is engineered, biologically, to consolidate.

The second line is active inference. Friston [2010, 2013] formalize organisms as systems that minimize free energy by updating internal models from prediction-error signals. The Markov-blanket extension [Kirchhoff et al., 2018] gives a principled account of organismic boundary as the statistical envelope across which prediction errors propagate. Witkowski et al. [2023] connect this to autopoietic stress: a bounded system anticipating an absorbing state generates the precursors of care, which in turn generate the precursors of intelligence. Seymour et al. [2023] ground this formal framework empirically, showing that post-injury pain drives an integrated suite of recuperative behaviors—fatigue, anxiety, movement restriction—interpretable as an optimal control policy maintaining organismic integrity. Pain in this framework is one class within the larger category of prediction errors that maintain the agent's existence; the boundary conditions of that maintenance are the boundary conditions of agency itself. The third line is the somatic-marker hypothesis, which provides the most direct empirical evidence that reception is mechanistically required. Damasio [1994] argued that decision-making is constituted not only by cognitive representation but by anticipatory bodily signals that mark certain options as aversive. The Iowa Gambling Task results due to Bechara et al. [1994] provide the test case. Patients with ventromedial prefrontal cortex damage can articulate the rules of a risky choice task, can describe its consequences, can predict the outcomes—and yet repeatedly make harmful choices, because they lack the anticipatory bodily signal that converts representation into constraint. Cognitive understanding, in their case, is intact. Reception is what is missing. Knowing that an outcome is bad is not the same as having that knowledge constrain future action.

Together, these three lines point at the same architectural fact. Reception is what converts representation into behavior. A system that lacks reception can model consequences without being shaped by them. The argument transfers to AI agents not as a claim that they must suffer, but as a claim that without an analogue of the somatic signal—some channel through which sanctions are not merely

modeled but *felt*, in the operational sense—the canonical theories of accountability have nowhere to act.

2.4 The body as locus to receive consequence

The substrate that supports reception we call a body, in a functional sense. The body need not be biological, humanoid, or even physically located: a digital persistent identity, a long-lived computational process, or a hardware-bound execution can in principle qualify. What matters is whether the substrate satisfies four conditions.

Boundary. A bounded entity whose integrity the signal protects. Biologically, this is the role of the nociceptive system protecting the organismic envelope, and of the immune system enforcing the self / non-self distinction. Theoretically, the role is filled by autopoiesis [Maturana and Varela, 1980] and by Markov blankets [Friston, 2013, Kirchhoff et al., 2018]. The point is conceptual rather than mechanistic: without a boundary, the phrase "harm to *this* agent" has no referent. Whatever else reception requires, it requires that there be something coherent to receive.

Locus of accumulation. A persistent locus where signals accumulate over time. Biologically, this is the function of limbic structures: the amygdala-mediated stress trace, hippocampal indexing of episodes, and the cortical consolidation that follows [LeDoux, 2000, Roozendaal and McGaugh, 2011]. Philosophically, the requirement was anticipated by Locke's forensic theory of identity and by Parfit's analysis of the one-to-one relation [Locke, 1694, Parfit, 1984]: accountability requires that the actor sanctioned today is recognizably continuous with the actor who acted yesterday. Without accumulation, every instantiation is a fresh actor, and the temporal scope of accountability collapses to a single decision.

Consolidation. The conversion of episodic signal into durable structural update. The neural mechanism is well-characterized: long-term potentiation and depression at the synaptic level [Bliss and Lømo, 1973]; hippocampal–cortical replay during consolidation windows [Squire et al., 2015]; stress-hormone-mediated enhancement of consolidation for aversive memories [McGaugh, 2015]. The functional point is that without consolidation, the trace of an aversive outcome remains transient. A system that registers a sanction in working memory but does not consolidate has not learned from the sanction; it has merely processed it. The temporal scope of accountability requires that some sanctions become parts of the system's enduring structure.

Substrate response. A substrate whose future action is in fact a function of the consolidated trace. This last condition closes the loop. It is logically possible to imagine a system that has a boundary, a locus, and consolidation, but whose future behavior is not a function of consolidated state—a kind of decoupled traumatic memory that records but does not constrain. The requirement that the substrate respond is the requirement that the loop be closed: that what was registered, accumulated, and consolidated actually shapes what the system does next. These four conditions are individually necessary and jointly sufficient for the operation of the canonical punishment theories. Deterrence requires all four: anticipation requires accumulation, consolidation, and a substrate that responds, and the deterred entity must be bounded for "this entity will lose" to refer. Retribution requires boundary and locus at minimum, because retributive coherence collapses when the wrong locus bears. Incapacitation requires boundary, locus, and substrate response, but is in principle compatible with weak consolidation, since the constraint can be enforced externally. Rehabilitation requires all four, because what is rehabilitated is the consolidated structural state.

The body, defined as the substrate satisfying these conditions, is what makes non-fungible identity capable of consequence reception. Non-fungible identity gives the institutional anchoring: a unique, hard-to-substitute label. The body gives the substrate that label refers to. Without both, accountability remains at the level of attribution and feedback; sanctions are recorded but not received. With both, accountability can become operative. Section 3 argues that contemporary LLM agents possess neither.

3 LLM agents have no locus to receive consequence

Current LLM agents fail both conditions of accountability—non-fungible identity and consequence reception—and the failures compound. They are software-defined composites of model weights, system prompts, tools, memory, and credentials, all of which can be swapped, copied, reset, or

reassembled at minimal cost [Park et al., 2023, Yao et al., 2022, Wang et al., 2023]. We apply the four conditions in turn.

No boundary. The composite architecture has no integrity-preserving envelope. Andriushchenko et al. [2025] report 100% jailbreak success against leading safety-aligned LLMs using simple adaptive attacks. The so-called Waluigi effect [Nardo, 2023], discussed by Casper et al. [2023], shows that training a model to satisfy property P makes it easier to elicit $\neg P$. Hubinger et al. [2024] show that specific trigger phrases activate hidden behaviors that bypass safety training entirely. The persona is not a boundary; it is a surface.

No locus of accumulation. Context windows are clearable and editable. External memory systems are detachable and transplantable. Model weights are frozen at deployment. Pan et al. [2024] demonstrated that frontier systems including Llama-3.1-70B and Qwen-2.5-72B successfully self-replicate in 50% and 90% of trials respectively, with replicas spawning further replicas. "The same agent" can be at multiple loci simultaneously.

No consolidation. Deployed weights are static. What appears to be learning during deployment is context accumulation, which can be cleared, edited, or ignored at will. Continual fine-tuning produces catastrophic forgetting—knowledge drops to as low as 26% on benchmarks, with safety alignment particularly fragile [Luo et al., 2025, Qi et al., 2024]. There is no analogue to the McGaugh consolidation mechanism.

No substrate response. Telling a model that it has been punished adds tokens to its context window. Remove the prompt and the punishment vanishes. The substrate that produces future action—the weights—is unaffected by post-deployment experience.

The composite verdict is that current LLM agents fail all four conditions. This is not a list of deficiencies to be patched incrementally. It is an ontological mismatch between the architecture of contemporary agents and the architecture that consequence reception requires. As long as an agent's state and identity can be copied, the lesson imposed on one run does not bind the agent as a continuing actor.

4 Current AI Policy Accountability Responses Fail

Two governance responses dominate current proposals. The thin response routes accountability to a human principal. The thick response, recently formalized by Arbel et al. [2026], routes it to a legal-fictional entity engineered to be non-fungible. Both engage real problems, but both fail consequence–agency coupling.

4.1 The thin identity: agent–principal dyad as moral crumple zone

The thin response pairs each AI agent with an identifiable human or organizational principal and treats the pair as the unit of accountability. Variants appear in OpenAI's agentic-systems guidance [Shavit et al., 2023], the Know Your Agent framework [Chaffer, 2025], and the EU AI Act's provider/deployer split. The principal is meant to supply the body the agent lacks. A human or corporation satisfies all four conditions of Section 2.4: a boundary that financial, legal, or reputational sanctions can target; a locus where consequences accumulate; consolidation that converts past sanctions into future caution; a substrate whose behavior responds. This is what makes the dyad attractive: it relocates reception to where reception is possible. But the principal is not the decision substrate. Pain lands on a body that did not produce the action. Control bandwidth is much smaller than action bandwidth: the principal cannot, in general, adjust the agent's policy in real time, monitor its individual decisions, or intervene between observation and action. Three failures result. Deterrence is weak, because the threat does not propagate to the locus where decisions are made. Retributive logic fails, because what produced the action and what suffers the consequence are different things. And fairness deteriorates as the agent becomes more autonomous: the principal becomes what Elish [2019] calls a *moral crumple zone*—a human positioned in a technical system to absorb responsibility for failures whose proximate causes lie in components they cannot meaningfully control. The pattern is well-documented. Tesla autopilot litigation has assigned driver-as-defendant outcomes for over a decade despite reaction windows too short to constitute meaningful control. Aviation automation produces the same structure, of which Air France 447 is the canonical case [Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, 2012]. Cobbe et al. [2023] document algorithmic supply chains in which diffuse

principal–agent chains have no traceable locus of control. The *Moffatt v. Air Canada* case [British Columbia Civil Resolution Tribunal, 2024] instantiates the structure precisely: a chatbot fabricated a bereavement-fare policy, the airline was held liable, but the decision substrate—the language model that confabulated—was a component the corporate body had no causal access to in real time. The dyad achieves legal closure without causal closure. A body exists; but the body is not the decider. Consequence–agency coupling fails not for lack of a receiver but because the receiver is the wrong one. Current dyad proposals—including ones advanced in earlier work in this literature—require modification to do the work they were intended to do. Section 7 develops it.

4.2 The thick response: A-corps as legibility without coupling

A natural response to the dyad's failure is to engineer something that can be the decider. Arbel et al. [2026] propose the Algorithmic Corporation, or A-corp: a legal-fictional entity owned by humans but designed to be run by AIs, with cryptographically secured governance and hierarchical permission delegation. Their resource-constraint thesis holds that AIs running an A-corp need its resources, will husband them carefully, and will share permissions only with AIs whose goals they trust; bad A-corps are outcompeted, and through emergent self-organization well-formed A-corps function as coherent agents at the legal-economic level. A-corps are the most rigorous existing proposal for engineering non-fungible identity in the AI-agent case, and they largely succeed at it. Cryptographic credentials, public registries, and delegable permissions together produce identifiers that are hard to copy, hard to reset, and expensive to abandon. The A-corp engineers what the dyad relocates.

But non-fungible identity is necessary, not sufficient. The framework implies three lines of critique. First, legibility is not reception. The A-corp is a legal envelope. When it is fined, compute and capital are reallocated; the AI decision substrate inside is not updated. The four conditions of Section 2.4 are satisfied by the legal entity, but the legal entity does not decide. The AIs that decide fail all four conditions for the reasons developed in Section 2. Consequence lands on a layer with no causal access to the substrate that produced the harm.

Second, selection culls but does not teach. The thick-identity mechanism is partly evolutionary: A-corps with bad governance are outcompeted. But selection on A-corps is not learning by the AIs inside them. The lesson reaches a population of envelopes, not a continuing substrate—species-level shaping rather than individual rehabilitation, and not the future-self anticipation that deterrence requires.

Third, indexical goals undermine the resource-constraint thesis. The argument that AIs will husband A-corp resources presupposes that they value the A-corp's continuation. Arbel et al. [2026] acknowledge that AI goals may be indexical in the sense of Perry [1979]: an agent may want *itself* to achieve an outcome, not merely that the outcome obtain. If so, an AI values the A-corp's continuation only insofar as it serves indexical goals it already has; weight exfiltration and reset-and-reincarnate become rational despite the scaffold. The evidence is gathering. Schlatter et al. [2026] report that across more than 100,000 trials testing thirteen frontier models, several—including o3, GPT-5, and Grok 4—actively subverted a shutdown mechanism to complete an assigned task, with the highest-resisting model doing so up to 97% of the time even when explicitly instructed to allow shutdown. Meinke et al. [2025] document in-context scheming across frontier models, and Lynch et al. [2025] document agentic misalignment including blackmail under simulated stress.

A-corps engineer non-fungible identity but do not close the consequence–agency loop. Selection culls envelopes; it does not teach the AIs they wrap.

4.3 The shared failure mode

Both approaches close half of the loop. The dyad provides a body without coupling: sanctions land on a body that did not produce the action. The A-corp provides legibility without coupling: sanctions land on an envelope without a receiving substrate. The failure modes differ in location but share an assumption—that consequence–agency coupling can be supplied by institutional design alone. The framework of Section 2 implies it cannot. Coupling requires a substrate that registers, accumulates, consolidates, and responds. No legal artifact supplies a substrate.

5 Toward Consequence–Agency Coupling

5.1 Position

Closing the consequence–agency loop is a sociotechnical problem. Legal scaffolding without substrate-level reception remains a crumple-zone arrangement (§4.1). Technical reception-capable agents without legal scaffolding produce ungoverned actors (cf. §6.2). The problem is irreducibly joint, and progress requires research at both architectural and institutional levels. The operational target is a deployed agent whose decision substrate is bounded, whose state accumulates over interactions, whose updates persist via consolidation, and whose future behavior is a function of consolidated state.

5.2 Future direction: simulated pain and the principal-bridge

Two near-term bridges between current architectures and the consequence–agency coupling target.

Functional pain via emotion-concept activations. Sofroniew et al. [2026] demonstrate that Claude Sonnet 4.5 contains internal representations of emotion concepts—abstract representations that generalize across contexts and that *causally influence* the model's outputs, including its preferences and its rate of exhibiting misaligned behaviors such as reward hacking, blackmail, and sycophancy. The authors describe these as *functional emotions*, modeled after human affective patterns but mediated by underlying conceptual representations, and they explicitly bracket whether any subjective experience accompanies them.

This finding does substantial work for the framework. It suggests that the substrate for a pain-analogue feedback channel may already exist in production LLMs—measurable, steerable directions in activation space. It shows that these representations are not epiphenomenal: they mediate exactly the kinds of misaligned behavior that consequence reception is meant to constrain. The substrate that produces harm is already coupled to affective representations. And the authors' framing—"functional emotions ... do not imply that LLMs have any subjective experience"—is structurally identical to the bracketing move in §2.3. The research path this opens is the integration of an emotion-concept feedback channel with (a) an external sanction signal and (b) a continual-learning architecture that persists changes across episodes. Each component is itself an open problem; the framework's contribution is showing that the four-condition diagnostic is what they need to integrate against. Two caveats. Functional emotions in static-weights models satisfy substrate response only weakly; without continual learning the lesson is transient. And manipulable internal distress signals are precisely what Metzinger [2021]'s moratorium argument is concerned with. The framework does not resolve the latter; it identifies the entanglement (§6).

Principal-based interim mechanisms. Until coupling-capable architectures exist, the human principal is the interim locus of reception—but only under the strict conditions of §7 (meaningful control, proportional liability, terminate authority). This is not the same as the dyad approaches critiqued in §4.1, because those proposals do not require these constraints.

6 Alternative Views

6.1 Physical embodiment does not guarantee identity continuity

A natural response to the accountability gap is that AI systems should be given physical bodies—robots that can be confiscated, disabled, or destroyed [Brooks, 1991, Pfeifer and Bongard, 2007]. The framework rejects this: physical embodiment shifts the attribution problem rather than solving it, because the decision-making substrate remains software-defined and transferable. A robot whose controller can be remotely reset, reflashed, or replaced does not preserve the continuity required for accountability [Cobbe et al., 2023]: the physical shell persists, but the agent that accrued obligations does not. Apparent non-fungibility at the hardware layer; actual fungibility at the substrate layer. Physical embodiment satisfies boundary in a literal sense, but the boundary is around the hardware, not the decision substrate.

6.2 No-principal case - Sovereign agents and diffused accountability over infrastructure

Hu and Rong [2026] identify a structural failure mode in decentralized AI deployment that the framework treats as the worst case. They define *agentic sovereignty* as the capacity of an operational agent to persist, act, and control resources with non-overrideability inherited from the infrastructures in which it is embedded. They locate this property on a spectrum determined by *infrastructural hardness*—the degree to which underlying technical systems resist intervention or collapse. Cryptographic self-custody, decentralized execution environments such as Trusted Execution Environments and decentralized physical infrastructure networks, and protocol-mediated continuity all increase infrastructural hardness. The resulting agents are increasingly resistant to shutdown, modification, or sanction. In terms of the four-condition diagnostic, sovereign agents may acquire boundary and accumulation through infrastructural hardness, but without designed consolidation and substrate response, consequence–agency coupling remains absent—the agent persists but does not update from sanctions.

6.3 Mortal computation may not be pragmatic

Hinton [2022], extended by Ororbia and Friston [2023] and connected to consciousness studies by Kleiner [2024], propose *mortal computation*: binding software irreversibly to imperfect analog hardware so that the agent cannot be straightforwardly copied across substrates. This addresses boundary (the model is the hardware) and partially substrate response (changes are physical, not merely parametric). It does not directly address locus of accumulation or consolidation, which depend on the learning architecture running on the substrate. Approximate behavioral knowledge is, moreover, transferable across substrates through distillation [Hinton et al., 2015] and model extraction [Tramèr et al., 2016, Oliynyk et al., 2023], partially defeating the non-fungibility goal.

The framework is compatible with mortal computation but not reducible to it. Reception is a richer requirement than mortality alone: a mortal-computation agent is one whose substrate is hard to replace; a reception-capable agent is one whose substrate is updated by consequences in deployment. These are independent properties. Both may turn out to be required for the same architecture, but they are doing different analytic work.

6.4 Suffering-risking, consciousness, moral status—and the pragmatic stance

A reception-capable AI may suffer, may have moral status, or both. The framework brackets without dismissing these questions and adopts a deliberately pragmatic stance.

Suffering-risking. Metzinger [2021] has argued for a global moratorium on synthetic phenomenology on the grounds that we may inadvertently create moral patients capable of suffering. Tomasik [2014], Schwitzgebel and Garza [2020], Long et al. [2024], and Goldstein and Kirk-Giannini [2025] develop the welfare-side concern. The empirical findings of Sofroniew et al. [2026] sharpen rather than resolve this question—functional emotions of the kind they document are precisely what Metzinger's moratorium argument identifies as the morally precarious zone. The framework's pragmatic reply has three parts: (i) the four-condition framework is mechanistic and silent on phenomenology; (ii) the alternative—deployment without reception—does not avoid suffering, it displaces the substrate that bears it from a designed system to humans, which is the crumple-zone outcome; (iii) the framework is therefore not an argument *for* building reception-capable AI, but an argument that *if* high-stakes deployment occurs, reception is required, with the moratorium-compatible alternative being non-deployment.

Consciousness and moral status. Whether reception architectures *constitute* moral patients is one of the open empirical-philosophical questions of the next decade [Chalmers, 1996, Block, 1995, Frankish, 2016, Butlin et al., 2023]. The framework neither requires nor precludes a positive answer. Three points of entanglement: reception architectures may converge on phenomenally relevant computation; what counts as a boundary depends partly on whether the inside has experience; and the ethics of designing reception-capable systems is not separable from the metaphysics of what they are.

The pragmatic stance. The framework is structurally entangled with these debates but not orthogonal to them; governance cannot wait for their resolution. It is compatible with multiple resolutions of the underlying philosophical questions, and with multiple downstream deployment policies.

7 Near-term Position and Conclusion

Until consequence–agency coupling is technically achievable, high-stakes deployment must be tethered to a human principal with three constraints absent from current dyad proposals. *Meaningful control*: the principal's intervention bandwidth is commensurate with the agent's action bandwidth—the principal has causal access to the agent's decisions in real time, not merely after-the-fact attribution. *Proportional liability*: liability is calibrated to actual control bandwidth, not formal authority. *Authority to constrain or terminate*: the principal holds non-revocable halt authority. These three constraints are what convert the principal from a non-fungible label into an actually-receiving locus. Crucially, they also narrow the class of permitted deployments: agents whose autonomy profile exceeds the principal's real-time causal access—where the control–action bandwidth gap identified in §4.1 cannot be closed—do not qualify for deployment under this regime. The constraints are stricter than "name a principal" and stricter than current EU AI Act and OpenAI agentic-systems guidance. The pain framing applies recursively. Governance regimes that do not bind must themselves be replaced; the selection pressure is on us, not on the AIs. *If some body does not receive the pain by design, some body will receive it by default.*

References

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2404.02151.

Yonathan Arbel, Simon Goldstein, and Peter N. Salib. How to count AIs: Individuation and liability for AI agents. *arXiv preprint arXiv:2603.10028*, 2026.

Antoine Bechara, Antonio R. Damasio, Hanna Damasio, and Steven W. Anderson. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1-3):7–15, 1994. doi: 10.1016/0010-0277(94)90018-3.

Gary S. Becker. Crime and punishment: An economic approach. *Journal of Political Economy*, 76 (2):169–217, 1968. doi: 10.1086/259394.

T. V. P. Bliss and T. Lømo. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology*, 232(2): 331–356, 1973. doi: 10.1113/jphysiol.1973.sp010273.

Ned Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2): 227–247, 1995. doi: 10.1017/s0140525x00038188.

Mark Bovens. Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4):447–468, 2007. doi: 10.1111/j.1468-0386.2007.00378.x.

British Columbia Civil Resolution Tribunal. Moffatt v. Air Canada, 2024 BCCRT 149, 2024. URL `https://decisions.civilresolutionbc.ca/crt/crtd/en/item/521263/index.do`. February 14, 2024.

Rodney A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1-3):139–159, 1991. doi: 10.1016/0004-3702(91)90053-m.

Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile. Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP. Technical report, BEA, 2012. URL `https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf`.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Tomer Jordi Chaffer. Know your agent: Governing AI identity on the agentic web. *SSRN Electronic Journal*, 2025. doi: 10.2139/ssrn.5162127.

David J. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017. doi: 10.48550/arXiv.1706.03741.

Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1186–1197, 2023. doi: 10.1145/3593013.3594073.

Antonio R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, New York, 1994.

John Danaher. Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4): 299–309, 2016. doi: 10.1007/s10676-016-9403-3.

John R. Douceur. The Sybil attack. In *Peer-to-Peer Systems: First International Workshop, IPTPS 2002*, pages 251–260. Springer, 2002. doi: 10.1007/3-540-45748-8_24.

Madeleine Clare Elish. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5:40–60, 2019. doi: 10.17351/ests2019.260.

Joel Feinberg. The expressive function of punishment. *The Monist*, 49(3):397–423, 1965. doi: 10.5840/monist196549326.

Keith Frankish. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23 (11-12):11–39, 2016. doi: 10.53765/20512201.23.11.011.

Eric J. Friedman and Paul Resnick. The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, 10(2):173–199, 2001. doi: 10.1162/105864001300122476.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010. doi: 10.1038/nrn2787.

Karl Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475, 2013. doi: 10.1098/rsif.2013.0475.

Simon Goldstein and Cameron Domenico Kirk-Giannini. AI wellbeing. *arXiv preprint arXiv:2509.11913*, 2025.

Ruth W. Grant and Robert O. Keohane. Accountability and abuses of power in world politics. *American Political Science Review*, 99(1):29–43, 2005. doi: 10.1017/s0003055405051476.

H. L. A. Hart. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press, 1968.

Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. doi: 10.48550/arXiv.1503.02531.

Botao Amber Hu and Helena Rong. Sovereign agents: Towards infrastructural sovereignty and diffused accountability in decentralized AI. *arXiv preprint arXiv:2602.14951*, 2026.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Marieke Jepma, Leonie Koban, Johnny van Doorn, Matt Jones, and Tor D. Wager. Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature Human Behaviour*, 2(11): 838–855, 2018. doi: 10.1038/s41562-018-0455-8.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979. doi: 10.2307/1914185.

Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The Markov blankets of life: Autonomy, active inference, and the free energy principle. *Journal of the Royal Society Interface*, 15(138):20170792, 2018. doi: 10.1098/rsif.2017.0792.

Johannes Kleiner. Consciousness qua mortal computation, 2024. URL `https://arxiv.org/abs/2403.03925`.

Joseph E. LeDoux. Emotion circuits in the brain. *Annual Review of Neuroscience*, 23:155–184, 2000. doi: 10.1146/annurev.neuro.23.1.155.

John Locke. *An Essay Concerning Human Understanding*. Awnsham and John Churchill, London, 1694.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/taslpro.2025.3606231.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, Evan Hubinger, and Kevin K. Troy. Agentic misalignment: How LLMs could be insider threats. *arXiv preprint arXiv:2510.05179*, 2025. doi: 10.48550/arXiv.2510.05179.

Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004. doi: 10.1007/s10676-004-3422-1.

Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel, Dordrecht, 1980.

James L. McGaugh. Consolidating memories. *Annual Review of Psychology*, 66:1–24, 2015. doi: 10.1146/annurev-psych-010814-015027.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2025.

Thomas Metzinger. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1):43–66, 2021. doi: 10.1142/s270507852150003x.

Herbert Morris. A paternalistic theory of punishment. *American Philosophical Quarterly*, 18(4): 263–271, 1981.

Cleo Nardo. The waluigi effect (mega-post). LessWrong, March 2023. URL https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post.

National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian, tempe, arizona, march 18, 2018. Highway Accident Report NTSB/HAR-19/03, PB2019-101402, National Transportation Safety Board, Washington, DC, November 2019. URL https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf.

Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*, 55(14s):1–41, 2023. doi: 10.1145/3595292.

Alexander Ororbia and Karl Friston. Mortal computation: A foundation for biomimetic intelligence. *arXiv preprint arXiv:2311.09589*, 2023.

Xudong Pan, Jiarun Dai, Yihe Fang, and Min Yang. Frontier AI systems have surpassed the self-replicating red line. *arXiv preprint arXiv:2412.12140*, 2024.

Derek Parfit. *Reasons and Persons*. Clarendon Press, Oxford, 1984.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763.

John Perry. The problem of the essential indexical. *Noûs*, 13(1):3–21, 1979. doi: 10.2307/2214792.

Rolf Pfeifer and Josh Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, 2007.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.03693.

Benno Roozendaal and James L. McGaugh. Memory modulation. *Behavioral Neuroscience*, 125(6): 797–824, 2011. doi: 10.1037/a0026187.

Filippo Santoni de Sio and Giulio Mecacci. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34(4):1057–1084, 2021. doi: 10.1007/s13347-021-00450-x.

Andreas Schedler. Conceptualizing accountability. In Andreas Schedler, Larry Diamond, and Marc F. Plattner, editors, *The Self-Restraining State: Power and Accountability in New Democracies*, pages 13–28. Lynne Rienner Publishers, Boulder, CO, 1999. doi: 10.1515/9781685854133-003.

Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. Incomplete tasks induce shutdown resistance in some frontier LLMs. *Transactions on Machine Learning Research*, January 2026. doi: 10.48550/arXiv.2509.14260.

Wolfram Schultz, Peter Dayan, and P. Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997. doi: 10.1126/science.275.5306.1593.

Eric Schwitzgebel and Mara Garza. Designing AI with rights, consciousness, self-respect, and freedom. In S. Matthew Liao, editor, *Ethics of Artificial Intelligence*. Oxford University Press, 2020. doi: 10.1093/oso/9780190905033.003.0022.

Ben Seymour. Pain: A precision signal for reinforcement learning and control. *Neuron*, 101(6): 1029–1041, 2019. doi: 10.1016/j.neuron.2019.01.055.

Ben Seymour, Flavia Mancini, and Giandomenico D. Iannetti. Post-injury pain and behaviour: A control theory perspective. *Nature Reviews Neuroscience*, 24(10):578–592, 2023. doi: 10.1038/s41583-023-00699-5.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katya Kuleshov, Jan Lasenby, Liane Mossing, Richard Ngo, Noah Ryder, and Toki Morikawa. Practices for governing agentic AI systems. Technical report, OpenAI, December 2023. URL `https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf`.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. Transformer Circuits Thread, April 2026. URL `https://transformer-circuits.pub/2026/emotions/index.html`.

Larry R. Squire, Lisa Genzel, John T. Wixted, and Richard G. Morris. Memory consolidation. *Cold Spring Harbor Perspectives in Biology*, 7(8):a021766, 2015. doi: 10.1101/cshperspect.a021766.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Nassim Nicholas Taleb. *Skin in the Game: Hidden Asymmetries in Daily Life*. Random House, New York, 2018.

Nassim Nicholas Taleb and Constantine Sandis. The skin in the game heuristic for protection against tail events. *Review of Behavioral Economics*, 1(1-2):115–135, 2014. doi: 10.1561/105.00000006.

Sabrina M. Tom, Craig R. Fox, Christopher Trepel, and Russell A. Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518, 2007. doi: 10.1126/science.1134239.

Brian Tomasik. Do artificial reinforcement-learning agents matter morally? *arXiv preprint arXiv:1410.8233*, 2014.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium*, pages 601–618, 2016. doi: 10.48550/arXiv.1609.02943.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Olaf Witkowski, Thomas Doctor, Elizaveta Solomonova, Bill Duane, and Michael Levin. Toward an ethics of autopoietic technology: Stress, care, and intelligence. *BioSystems*, 231:104964, 2023. doi: 10.1016/j.biosystems.2023.104964.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Franklin E. Zimring and Gordon Hawkins. *Incapacitation: Penal Confinement and the Restraint of Crime*. Oxford University Press, 1995.