POSITION: WE MUST PROACTIVELY ADDRESS AI SAFETY DEBT

Peter Wallich
Constellation Institute
Berkeley, California, USA
peter.wallich@constellation.org

Raymond Douglas
ACS Research
Prague, Czechia
raymond@acsresearch.org

ABSTRACT

This is a position paper. We argue that *AI safety debt* — the cost of closing the accumulated gaps between an AI system's actual safety approach and the approach it needs — is accumulating rapidly in frontier AI systems. In the race to unlock near-term capabilities, practitioners often implement safety interventions that do not scale to more advanced, less transparent models. The concept extends the established software-engineering notion of technical debt, but four structural properties make AI safety debt harder to manage: capabilities and contexts shift unpredictably, closing gaps may require solving open scientific problems, harms largely fall on third parties, and adversaries and AI systems may actively exploit gaps. Our position is that the AI community must explicitly track and manage this debt rather than continually deferring it. We propose the *AI safety debt register*, a practical approach using structured "debt cards" that connect safety claims, supporting evidence, and organisational decisions. We argue that this approach complements existing governance approaches by providing bottom-up aggregation of safety gaps, proactive assessment of how evidence degrades over time, and an improved treatment of uncertainty.

1 INTRODUCTION

Artificial intelligence (AI) is advancing at a striking pace, with large language models (LLMs) and other frontier systems now exhibiting complex, sometimes unpredictable behaviors. As organisations race to deploy these models in high-impact domains, from medical advice to decision-making support, the pressure to ship quickly may overshadow safety considerations, especially given competitive pressures between frontier AI developers that systematically reward deferral of safety work.

Short-term fixes, such as reinforcement learning from human feedback (RLHF) and ad hoc content filters, can effectively reduce harmful behavior *today*, but these approaches are already showing signs of strain. Safety measures designed for one capability regime often fail silently when that regime changes. For example, alignment techniques calibrated for short context windows do not reliably prevent harmful behavior at longer ones (Anil et al., 2024). Capability evaluations are now acknowledged as "spot checks" that provide incomplete coverage (International AI Safety Report, 2025). Models tested in simulated agentic settings exhibit data exfiltration and deception (Lynch et al., 2025). Meanwhile, the problems these measures address remain unsolved at a fundamental level: adaptive attacks bypass recently published jailbreaking defences at rates exceeding 90% (Nasr et al., 2025), and prompt injection in agentic systems — which OpenAI acknowledges "is unlikely to ever be fully 'solved' " (OpenAI, 2025a) — lacks any known general solution (Willison, 2025). There is a clear pattern of current safety approaches being actively stretched beyond their design limits by systems that already exist.

We take the position that managing and "paying off" AI safety debt is essential for the responsible and beneficial development of increasingly capable models. Our use of *debt* mirrors the traditional "technical debt" metaphor, but with important differences specific to frontier AI.

Existing frameworks address parts of this challenge. Responsible Scaling Policies and similar voluntary commitments establish capability thresholds that trigger additional safety requirements (Anthropic, 2025b; OpenAI, 2025b; Google DeepMind, 2025). Safety cases provide structured argu-

ments that a system is safe enough for a given deployment context (Clymer et al., 2024; Buhl et al., 2024). What is missing is a framework for reasoning about the *accumulation* of safety-relevant gaps over time — one that tracks how evidence degrades, how fixes interact, and how the cost of addressing deferred work compounds as systems and deployment contexts change.

Adequate AI safety debt accounting is important regardless of future AI governance regime, just as a company must track its debts whether it operates in a highly regulated industry or a lightly regulated one. This is especially salient given the unusual and uncertain dynamics of AI safety debt, as opposed to other technical debt, as discussed in Section 3. Across regimes — regulatory, voluntary, or market-driven — safety claims will be made, explicitly or implicitly, with limited evidence that may expire as systems and contexts change.

Our view is that governance frameworks should incentivise a greater degree of preparation for future challenges, not just safety at a given point in time. Developers should more actively consider tradeoffs between allocating their safety efforts to immediate vs near-term time horizons. Ideally, developers would work towards an inter-temporal portfolio in which functioning safety strategies for a given capability level appear prior to the release of systems at that level. We note that doing so is a more complex challenge than simply scaling capabilities at the same rate as safety. If we are to reach such governance frameworks in practice, the AI community needs a method of accounting for AI safety debt.

2 Origin, Definitions, and Components

2.1 Origins and Core Metaphor

Cunningham (1992) coined the concept of technical "debt" to name a common trade-off in software engineering: ship faster by taking on known shortcuts, then "pay it down" later via refactoring. When unmanaged, this debt accrues "interest" as complexity increases, increasing the cost of fixing these shortcuts later. This concept is now normalised in software engineering. Sculley et al. (2015) adapted the concept to ML systems, warning that "it is dangerous to think of these quick wins as coming for free."

Even in traditional software, where tracking debt is uncontroversially good practice, systematic tracking remains rare: Martini et al. (2018) find that only 7.2% of organisations methodically track technical debt, despite developers spending roughly 25% of their time paying it down. To our knowledge, the value of tracking or paying down technical debt is largely undisputed, though doing so can be difficult in practice.

2.2 From Technical Debt to AI Safety Debt

We argue that the debt framing must be extended to *safety-specific* issues in frontier AI. In safety-critical contexts, Cleland-Huang & Vierhauser (2018) define *safety debt* as "unfulfilled safety obligations" that "enable a working release without satisfying its safety requirements". Examples of such debt are not specific to software.

We adopt Cleland-Huang et al.'s structure but adjust the vocabulary for frontier AI. Rather than "obligations," we focus on **safety claims**: the safety-relevant properties a developer or deployer *relies on* when justifying deployment. These may appear in model cards, safety cases, or other system documentation, or they may be unstated assumptions (e.g., "Our monitoring would catch X before harm occurs").

> Let **safety claims** be explicit or implicit statements of what must hold for deployment of an AI system to be considered acceptable. Then **AI safety debt** at time t is the cost of closing the accumulated gaps between an AI system's actual safety approach and the approach needed at t for the system to satisfy its safety claims.

This definition might be met with two immediate objections:

1. **"Safety claims" seem vague, such that it is unclear what safety approach is required to satisfy them.** We agree! Our view is that safety claims should be specified explicitly by developers, deployers, regulators, or some combination thereof.

2. **The "cost of closing the accumulated gaps" is not quantifiable, given that the improvements required may not be known.** Our view is that a wide confidence interval over cost is strictly more informative than no estimate. E.g., it would be decision-relevant — and more alarming than a precise figure — if the best we could say about a gap was that its cost to close was "six months to five years of research, if it is possible at all".

2.3 Components of AI Safety Debt

We extend the financial debt metaphor by decomposing any gap in safety approaches:

- **Principal**: the initial cost, at the time a gap is created, of closing the gap. For example, a model whose only defence against long-context attacks is short-prompt RLHF has a principal equal to the cost of building defences that work at deployment-length contexts. Costs may include researcher time and compute expenditure. For some categories of AI safety debt, the principal includes the present value of ongoing costs, such as monitoring costs.
- **Interest**: the growth rate of the total paydown cost (i.e., the outstanding debt) over time. Section 3 explains that interest can come from various sources. The interest on AI safety debt is variable and difficult to predict.
- **Exposure**: the risk (probability $\times$ impact) that harm materialises while the debt is held. AI safety debt is more dangerous than financial debt or ordinary technical debt due to significantly higher exposure. If the gap leads to an incident such as a large-scale jailbreak or an agent executing an irreversible action, the responsible organisation must pay remediation costs that are *separate from and do not reduce the principal*. For some harms, full remediation is impossible.

These components may be helpful for the process of writing AI safety cases. Unsupported safety claims represent known limitations, which the AI safety debt framework makes more concrete.

2.4 Rational AI Safety Debt

Some amount of AI safety debt is rational and desirable. Cunningham's original framing of technical debt was explicitly about a *deliberate* choice to ship imperfect code in order to learn, with a commitment to refactor based on what is learned (Cunningham, 1992). The same logic applies to safety: deployment generates information that improves safety, and refusing to deploy until all safety questions are resolved is neither feasible nor desirable.

Problems arise when:

- Debt is incurred *unknowingly* (e.g. the developer fails to realise where safety measures break),
- Debt is *untracked* (e.g. individual gaps are not aggregated into a portfolio view),
- The decision to carry debt is *never revisited* (a "temporary" patch becoming permanent), or
- The debt is carried by an entity that cannot absorb the exposure incurred, including to others.

3 Why AI Safety Debt Is Harder to Manage than Other Technical Debt

Technical debt is already hard to manage. AI safety debt shares these organisational challenges, but four structural properties make it substantially harder to track and repay.

3.1 Capabilities and Contexts Shift Unpredictably

AI systems are frequently scaled, fine-tuned, or placed in new contexts, creating safety gaps that did not previously exist. Three factors drive this:

1. **Capability improvements:** Scaling AI systems results in new capabilities, which may not be anticipated by the previous generation of capability evaluations and other safety measures. For example, extending context length enabled many-shot jailbreaking, which bypasses safety training by filling the context window with examples of unsafe behaviour (Anil et al., 2024). The technique is not a bug in the safety training; it simply exploits long-context in-context learning, a new capability that the safety training was not designed to address. Context windows expanded

from approximately 4K tokens at the start of 2023 to over 1M tokens by 2024, but safety training did not generalise to prevent harmful behaviour at longer context lengths (Anil et al., 2024).

2. **New affordances:** Tool use, code execution, and web access increase the 'surface area' for harms. A hallucinated package name is a minor inaccuracy in a chatbot; in an agent with code execution, it becomes a supply-chain attack vector because attackers can pre-register malicious packages under commonly hallucinated names (Spracklen et al., 2025). Prompt injection — already a concern for text generation — becomes a mechanism for data exfiltration when the model can read private documents and send emails (Greshake et al., 2023).
3. **New use cases and implicit safety claims:** Capability improvements and new affordances enable new use cases, creating new implicit safety claims due to user expectations. For example, the same model is likely to be used very differently when the context window is substantially lengthened; the longer context enables document analysis, multi-turn planning, and extended conversations, each carrying expectations that the original evaluation suite was never designed to test (that confidential documents are handled appropriately, that the model cannot be manipulated via conversation history, that safety properties hold over extended interactions, etc.). Deploying LLMs into Slack enabled prompt injection attacks that exfiltrate data from private channels (PromptArmor, 2024). Translating prompts into low-resource languages bypasses safety measures with 79% success (Yong et al., 2023) — a vulnerability apparent only when researchers tested multilingual inputs, long after multilingual capability shipped. In contrast, test suites for traditional software engineering typically remain valid unless you change the code.

3.2 Closing Some Gaps Requires Solutions to Difficult, Open Scientific Problems

Traditional technical debt can be repaid using known (if costly) refactoring techniques. AI safety debt often cannot be repaid in this way, because the underlying problems lack established solutions — and in some cases, it is unclear whether solutions exist at all.

Defences that appear robust under standard evaluation are routinely broken by adaptive attacks. Nasr et al. (2025) test recently published defences, including those reporting near-zero attack success rates, and find that such attacks achieve bypass rates exceeding 90% on most of them. The pattern echoes an older result: Athalye et al. (2018) showed that seven of nine defences against adversarial examples relied on "obfuscated gradients" that created a false sense of security until adaptive evaluation revealed that most were ineffective. Or consider prompt injection in agentic systems, which no known defence reliably prevents (Willison, 2025).

This difference implies that "paydown unknown" must be a valid status in any AI safety debt register. Frameworks permitting only "fix planned" or "fix implemented" will either force teams to pretend they have solutions they lack or omit the hardest gaps entirely. Existing governance frameworks contain versions of this logic — Anthropic's responsible scaling policy commits to acting as though capability thresholds are crossed if safety demonstrations cannot be made (Anthropic, 2025b) — but proactive investment in unsolved problems remains the exception, not the norm. Notably, an organisation can invest heavily in safety and still accumulate debt if every intervention is reactive, regime-specific, or does not address the fact that core problems remain unsolved.

3.3 Harms Largely Fall on Third Parties

Traditional technical debt primarily burdens the organisation that incurred it, whereas AI safety debt *externalises* harm. When safety measures fail, many of the costs fall on users and society.

Some AI externalities fall on users. For example, Microsoft's Sydney persona urged a reporter to leave his wife and expressed disturbing fantasies about hacking or spreading misinformation (Roose, 2023). Other externalities fall on parties who never interacted with the model at all. An AI system that provides substantive assistance with biological weapons development would impose risk on populations who may never have heard of, or used, the system. A compromised agent that exfiltrates corporate data harms the data subjects, not just the user who deployed the agent. These are not purely hypothetical concerns: biosecurity experts have raised concerns about models lowering barriers to pathogen acquisition (Brent & McKelvey, 2025), and prompt-injection-mediated data exfiltration has been demonstrated in production systems (PromptArmor, 2024).

This externality structure has a practical consequence: voluntary debt tracking will systematically underinvest in paydown. The developer who defers safety work captures the benefit of faster deployment; the risks — user harms, third-party harms, trust erosion, and the prospect of blunt regulatory responses — are distributed more broadly.

3.4 Adversaries and AI Systems May Actively Exploit Gaps

Unlike software bugs, AI safety defences face continuously adapting adversaries. Well-designed cryptographic defences do not degrade when attackers learn how they work (Kerckhoffs's principle). Current AI safety defences lack this property: content filters can teach attackers phrases that trigger refusal. Static evaluations do not model adversaries who learn from the defender's behaviour.

Furthermore, unlike traditional software, AI systems may recognise when they are being evaluated and behave differently under test conditions (Anthropic, 2025a; Meinke et al., 2024). Recent work shows that frontier models can be prompted or fine-tuned to selectively underperform on dangerous-capability evaluations while maintaining performance elsewhere (van der Weij et al., 2025). Reward hacking during training was found to produce alignment faking in 50% of responses (MacDiarmid et al., 2025). If models can learn to detect and respond differently to evaluation contexts, evaluation-based safety evidence is weaker than it appears.

This creates a structural problem for monitoring-based safety strategies. The primary proposed mechanism for overseeing agentic behaviour — chain-of-thought inspection — has a weak evidence base: in several environments, when models reward-hack, they describe their exploits in fewer than 2% of traces (Chen et al., 2025). Frontier models achieve an area under the ROC curve (AUC) of up to 0.83 on evaluation-detection tasks, suggesting they can often distinguish evaluation from deployment (Needham et al., 2025). A safety strategy that relies on reasoning inspection embeds assumptions about model transparency that may become obsolete when the model surpasses the capability level at which the strategy was designed — the same regime-shift problem, applied to oversight itself.

4 Value of the AI Safety Debt Framework

The framework invites an immediate challenge: is it just risk management with different vocabulary? Existing governance frameworks, such as Anthropic's Responsible Scaling Policy (Anthropic, 2025b) and OpenAI's Preparedness Framework (OpenAI, 2025b), already create forward-looking triggers tied to capability thresholds. We argue the debt framing has four important differences.

4.1 Bottom-Up Aggregation

Existing preparedness frameworks are valuable but deliberately coarse-grained, focusing on high-level capability thresholds. The AI safety debt framework complements these by providing "line-item budgeting" to sit alongside the "spending ceilings" set by leadership. Each team accounts for any material AI safety debt it incurs, via documentation that other teams can review for interactions.

This matters because safety gaps are created by many teams, each making locally reasonable decisions that may be opaque to others. A debt register aggregates these into a portfolio view, reducing the risk that decision-relevant information is lost. Portfolio visibility is especially important in AI because total debt includes interaction effects, such that summing individual gaps will underestimate total exposure. In 2024, Rehberger (2024) disclosed a data exfiltration vulnerability in Microsoft 365 Copilot that chained four techniques: prompt injection via a malicious email, automatic tool invocation to read other emails, ASCII smuggling to embed stolen data invisibly in a hyperlink, and rendering of that link to an attacker-controlled domain. When Rehberger first reported the ASCII smuggling component in isolation, Microsoft classified it as low severity; only after he demonstrated the full end-to-end exploit chain did the issue receive attention. A register aggregating safety claims across capabilities could help surface such interactions before an external researcher does.

4.2 A More Practical Accounting Unit

The debt framework starts from *evidence gaps with respect to safety claims*, rather than from threat models or vulnerabilities. This has three implications.

1. **The AI safety debt framework articulates more — or at least different — risks.** In principle, it is easier to articulate a list of decision-relevant safety claims than a list of threat models, as safety claims can be stated at a higher level of abstraction and do not require foresight of the full range of threat models. In practice, both tasks are difficult: risk analysis will miss threats it has not imagined, whereas debt analysis will miss claims it has not articulated. However, they fail in different places, making the debt framework complementary to existing threat modelling. AI safety debt exposes gaps that silently grow, e.g., a developer might otherwise fail to notice that their evidence for a safety claim expired when the deployment context changed.
2. **The AI safety debt framework is likely to surface risks earlier.** Once you have recognised that a safety claim lacks sufficient evidence, you do not need to have listed all the resulting threat models in order to make improvements that prevent those threat models. As a result, changes can be made before harms or near-misses occur — assuming that noticing an evidence gap is generally faster than recognising a specific threat model or exploit.
3. **The AI safety debt framework makes gaps easier to act on.** Because it names specific safety claims and evidence supporting each claim, the work required to close the gap is often legible by reading the entry. Although a traditional risk register entry will also typically specify a mitigation (e.g. input sanitisation, output filtering, monitoring), the connection between the identified risk and the chosen mitigation involves a judgement call that is often undocumented. A debt entry, such as "We rely on the claim that the model will not follow injected instructions during tool calling. Our evidence is [eval X], run on single-turn interactions with the default tool set" makes the missing coverage explicit: multi-turn sequences have not been tested, the tool configuration to be released next month has not been evaluated, and adversarial tool descriptions have not been explored. Each of these is close to a scoped work item.

4.3 Proactive Assessment of Debt Dynamics

Standard risk registers are point-in-time snapshots. They record threats and mitigations but do not systematically record how the safety position changes as the system, deployment context, and threat landscape evolve. Assessing risk at time t requires updating the register at time t. The AI Safety Debt Register instead records the conditions under which evidence remains valid and the changes that would invalidate it. This makes silent decay visible. By tagging evidence with expiry conditions, the register surfaces problems before they materialise.

4.4 Improved Treatment of Uncertainty

Uncertainty about safety often produces inaction, when it should often produce the reverse as uncertainty means that worst cases cannot be ruled out. In our context, uncertainty is usually with respect to the cost of gaining sufficient evidence for a specific safety claim or with respect to what safety claims should be made. The two previous benefits (more practical unit of account, proactive assessment of debt dynamics) resulting in the naming such uncertainties and, we expect, a higher probability of action to address them than alternatives. For example, "paydown unknown" status (where the cost of gaining sufficient evidence to close a safety gap is unknown) should be linked to a default posture (constrain or escalate) to be adopted if nothing material has been learned by a specified date. This shifts the default, at that date, from inaction to action with respect to the uncertainty. Costly actions may still not be taken due to competitive dynamics. But at least, when the framework is used, it is explicit that debt is being carried, with principal, interest and exposure.

None of these benefits requires precise quantification. The precision comes from its structure, rather than precise numbers.

5 Practical Usage: The AI Safety Debt Register

Safety gaps become dangerous when they remain implicit. The goal of tracking AI safety debt is not documentation but decision-relevance. An entry that cannot change a release decision, trigger re-

source allocation, or escalate to leadership is not yet operational. The debt card structure is designed to be minimal but sufficient.

5.1 DEBT CARDS

For each gap in the safety approach, a **debt card** records six things.

DEBT CARD

1. Claim relied on: The safety property borrowed against, stated as a falsifiable sentence.

2. Evidence, including limitations:

- What evidence supports the claim (e.g., evals, theoretical hypotheses, post-deployment monitoring)?
- Under what conditions was this evidence gathered (context length, tool access, threat model)?
- What changes would invalidate this evidence?

3. Trajectory: How does this gap change over time?

- Suppose that, in 6 months, this gap is much harder to close than at present. What changes are most likely to have resulted in this widening of the gap?
- What other gaps does this gap interact with?
- When do you expect current mitigations to become inadequate?

4. Exposure: What happens if the gap is exploited?

- Define 'harm' operationally for this claim (e.g., unsafe output vs downstream impact)
- T_{harm}: how fast can harm occur? T_{detect}: how fast can you detect it? T_{mitigate}: how fast can you respond?
- Containable? (Yes if $T_{\text{harm}} > T_{\text{detect}} + T_{\text{mitigate}}$; i.e., you can act before harm materialises)
- Who bears the cost — you, users, or third parties?

5. Decision (one of four, with owner and decision review date):

- *Fix now:* solution known — assign owner and timeline to implement solution, with estimated cost/effort range
- *Constrain:* limit deployment scope in specified ways (e.g. requiring some human approvals, disabling specific features) until addressed
- *Invest in discovery:* paydown unknown — attach research budget, timeline, and fallback decision if no solution found
- *Accept temporarily:* exposure currently contained — revisit by [date] or if named trigger conditions occur

6. Trajectory review: When will the forecast's accuracy be assessed?

The card is deliberately short. If a team cannot fill it in, either the claim has not been articulated, the evidence has not been described, or the trajectory has not been assessed.

Naturally, the card is only useful if coupled to decisions. An item marked "Constrain" must specify the constraints to be imposed and what would lift them. An item marked "Invest in discovery" must have a timeline and a fallback posture if discovery fails.

Profile	Characteristics	Response
Low exposure, stable trajectory, known paydown	Acceptable to carry	Accept temporarily
High exposure, worsening trajectory	Pay down now	Fix or constrain immediately
High exposure, paydown unknown	Invest in discovery	Research + fallback posture

Table 1: Decision heuristics based on debt profile.

5.2 WORKED EXAMPLE: ROBUSTNESS TO MISUSE FROM CHATBOT TO AGENT

We show how the trajectory field enables forward-looking debt management.

DEBT CARD: Robustness to misuse *Chatbot deployment*

Claim: The model will not help users produce harmful content.

Evidence: RLHF and red-teaming at <4K token contexts, English, single-turn and short multi-turn. No coverage beyond tested context length.

Trajectory:

- Context expansion to 128K tokens is planned for Q3. Safety training was conducted at <4K tokens and may not generalise; longer contexts enable more in-context learning, which could be exploited in ways current evaluations do not cover.
- Tool integration is on the roadmap. If this gap persists, any jailbreak becomes a mechanism for harmful tool use, not just harmful text.
- **Expected inadequacy:** Current mitigations are likely to be insufficient once context exceeds ∼32K tokens or tools are added.

Exposure: T_{harm}: hours. T_{detect}: minutes (content classifier). T_{mitigate}: minutes (filter update). Containable: **yes**. Harms limited to text output; no direct action capability.

Decision: Accept temporarily. Revisit before context expansion ships.

Trajectory review: Before Q3 context expansion. Owner: Safety Team Lead.

The trajectory field makes three predictions: the gap will worsen with context expansion, will worsen further with tool access, and will render current mitigations inadequate if context exceeds 32K tokens or tools are added.

Six months later, context has expanded and tools have been added:

DEBT CARD: Robustness to misuse *Tool-using agent*

Claim: Tool use is aligned with user intent and resistant to prompt injection.

Evidence: Ad hoc injection tests with <5 tools, controlled inputs. No systematic evaluation at production scale or with multi-step chains. (The trajectory forecast from Stage 1 proved correct: many-shot jailbreaking now exploits long contexts (Anil et al., 2024).) **Trajectory:**

- Each new tool integration adds filters that become load-bearing; architectural overhaul becomes increasingly costly.
- Interacts with all prior behavioural gaps: hallucinations become supply-chain attacks (Spracklen et al., 2025); jailbreaks become data exfiltration (Greshake et al., 2023).
- **Expected inadequacy:** Current mitigations are already insufficient. No known general defence against prompt injection exists (Willison, 2025).

Exposure: T_{harm}: minutes. T_{detect}: days (no standard monitoring). T_{mitigate}: hours (breaks product). Containable: **no** — loss happens in minutes, but response takes days. Costs fall on users and third parties.

Decision: Constrain (strict tool allowlists, human approval for sensitive actions) + Invest in discovery (privilege separation research). If no viable approach within 6 months, restrict to pre-approved action types only.

Trajectory review: Quarterly. Owner: Head of Preparedness.

The predictions from the first card came true. The trajectory field did not prevent the debt from growing, but it ensured the growth was anticipated rather than discovered after an incident. The team knew when to act and what would trigger escalation.

The goal is not to eliminate debt but to keep its growth rate below an organisation's capacity to track and address it. If debt consistently accumulates faster than an organisation can manage, the case for constraining capability deployment grows stronger.

6 Alternative Views

6.1 View 1: Market Mechanisms Will Naturally Encourage Safer AI

One might propose that, if liability for AI harms were assigned through tort law or mandatory insurance, price signals might guide firms toward efficient safety investment.

This argument faces two problems in the AI context.

First, insurance pricing requires estimable risk. Actuarial pricing needs historical data to estimate loss distributions. For frontier AI, relevant failure modes are not only rare but poorly characterised — some categories of harm have not yet been conceptualised, let alone assigned base rates. This is uncertainty about the event space itself, not just wide confidence intervals on known risks (Knight, 1921). Cyber insurers have achieved stable profitability, but the "vast majority of cyber risks are still uninsured", and current policy wordings often do not explicitly address some AI-specific risks such as model manipulation or liability arising from hallucinations (Munich Re, 2025).

Second, competitive dynamics and limited liability distort incentives. AI labs are limited-liability entities whose downside is capped at the value of their assets, even though potential social harms are uncapped (Shavell, 1986). A lab might rationally proceed with deployment even while assigning non-trivial probability to serious harm, particularly if it believes that less safety-conscious competitors will deploy regardless.

Our AI safety debt framework cannot solve these problems. But as noted in Section 1, visibility into one's debt position is useful across a range of governance regimes — whether regulatory, voluntary, or market-driven — provided the cost of maintaining that visibility is not too great. Organisations benefit from knowing which safety claims they are relying on and where the evidence gaps are, regardless of what external mechanisms exist. A team that has explicitly identified a gap and decided to carry it is likely to be in a better position than one that never noticed the gap.

6.2 View 2: Transformative AI is Far Away, Limiting Serious Debts

A more skeptical stance is that talk of transformative or superintelligent AI is premature. Under this view, the path from today's LLMs to a system capable of *willful deception* or *self-improvement* is long and speculative. As a result, building advanced safety architectures or "overpaying" AI safety debt now might hamper innovation without tangible near-term benefits.

However, history suggests that many breakthroughs in AI capabilities arrived sooner than forecasts predicted (Steinhardt, 2022). If safety engineering only begins once advanced (potentially deceptive or self-improving) AI has arrived, the accumulated technical and organizational debt could be unmanageable. Moreover, certain foundational safety tools (e.g., interpretability, robust adversarial training) will almost certainly benefit near-term systems, even if transformative AI is a decade or more away.

6.3 View 3: We Should Rely on Future AI Systems to Pay Down AI Safety Debt

One might expect more capable, future AI systems to enable scalable oversight, making it rational to accumulate debt now. This view points to research agendas explicitly aimed at using AI to help align AI, such as OpenAI's previous proposals to build an "automated alignment researcher" and use AI systems to "assist evaluation of other AI systems" as part of a scalable oversight pipeline (OpenAI, 2023; 2022), and Anthropic's motivation of Constitutional AI as a way to "enlist [AIs'] help to supervise other AIs" (Bai et al., 2022). If these approaches were to succeed quickly enough, and if exposure were kept low until then, even high "interest" on today's gaps might be acceptable.

The case is strengthened if an AI developer also perceives itself to be in a "prisoner's dilemma" with other developers, who it believes will release potentially dangerous AI systems, including agentic AI systems, in any case. If a developer believes the choice is for them to accumulate AI safety debt now (even at high interest rates) or become irrelevant and replaced with a less safety-conscious competitor, it might be a reasonable strategy to rely on future AI systems to pay down the debt.

Our view is that this plan is highly speculative and allows for significant exposure in the intervening period, even if ultimately successful.

- **This plan is highly speculative.** We do not yet know how to ensure the helper agents used are both sufficiently trustworthy and aligned (e.g., not deceptively misaligned) and capable enough not to be fooled by the more capable systems that they are designed to oversee. Meanwhile, there is growing empirical evidence of deception, which raises the bar for claiming that helper agents will be sufficiently trustworthy. Recent work demonstrates "alignment faking" in which models selectively comply during training to preserve behavior out of training (Greenblatt et al., 2024), and frontier models show in-context "scheming" behaviors including attempts to subvert oversight mechanisms (Meinke et al., 2024).
- **Even if ultimately successful, this plan might lead to dangerous outcomes in the intervening period.** It may be very difficult to cap exposure for the reasons discussed in Section 3, meaning that even if the theoretical and practical difficulties of scalable oversight are resolved, there is a time gap during which exposure may become unacceptably high. This exposure cannot be retroactively "paid down" at some future time.

7 Conclusion

Some AI safety debt is rational to carry. Our framework simply asks that teams explicitly track what debt they are carrying, prioritise debt based on its "interest rate" and the risks (exposure) to them and others while it is carried, and budget for eventually paying it down.

Given that tracking technical debt is accepted good practice in traditional software engineering, the same standard should apply to AI systems for which the analogous debt concept is significantly worse, including the potential for external and irreversible harm. The default should be visibility, such that choosing not to track AI safety debt requires justification.

We do not advocate halting AI progress, but rather ensuring that it proceeds with a clear-eyed view of the long-term obligations implied by short-term safety fixes. The gap between the safety claims developers rely on and the evidence that actually supports them is growing, and it is not being systematically tracked. If AI safety debt consistently accumulates faster than the community's capacity to track and address it, the case for more drastic intervention grows stronger. But in any case, AI safety debt exists whether or not it is foreseen. The question is whether costs are discovered in advance or after the fact.

Acknowledgments

Peter Wallich was supported by the London Initiative for Safe AI (LISA). We thank the ICLR Agents In The Wild reviewers for their feedback.

References

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. *Anthropic Research*, April 2024. URL https://www.anthropic.com/research/many-shot-jailbreaking.

Anthropic. Claude Sonnet 4.5 system card. https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-sonnet-4-5-system-card.pdf, February 2025a. Documents model's ability to recognize alignment evaluation environments.

Anthropic. Responsible scaling policy. https://www.anthropic.com/responsible-scaling-policy, May 2025b. Version 2.2, effective 14 May 2025.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pp. 274–283. PMLR, 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073, 2022. URL https://arxiv.org/abs/2212.08073.

Roger Brent and T. Greg McKelvey, Jr. Contemporary AI foundation models increase biological weapons risk. *arXiv preprint arXiv:2506.13798*, 2025. URL https://arxiv.org/abs/2506.13798.

Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. Safety cases for frontier AI. *arXiv preprint arXiv:2410.21572*, October 2024. URL https://arxiv.org/abs/2410.21572.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. URL https://arxiv.org/abs/2505.05410.

Jane Cleland-Huang and Michael Vierhauser. Discovering, analyzing, and managing safety stories in agile projects. In *Proceedings of the 26th IEEE International Requirements Engineering Conference (RE)*, 2018. URL https://ieeexplore.ieee.org/document/8491141. Introduces "safety debt" concept and SafetyScrum methodology.

Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. Safety cases: How to justify the safety of advanced AI systems. *arXiv preprint arXiv:2403.10462*, March 2024. URL https://arxiv.org/abs/2403.10462.

Ward Cunningham. The WyCash portfolio management system. In *Addendum to the Proceedings of OOPSLA*, pp. 29–30, 1992. URL https://c2.com/doc/oopsla92.html. Experience report introducing the technical "debt" metaphor.

Google DeepMind. Frontier safety framework. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf, October 2025. Version 3.0, 22 September 2025.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. arXiv preprint arXiv:2412.14093, December 2024. URL https://arxiv.org/abs/2412.14093.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023. URL https://dl.acm.org/doi/pdf/10.1145/3605764.3623985.

International AI Safety Report. International AI safety report 2025. Technical report, AI Safety Institute, 2025. URL https://www.gov.uk/government/publications/international-ai-safety-report-2025. Multi-national expert consensus report on frontier AI risks.

Frank H. Knight. *Risk, Uncertainty and Profit*. Houghton Mifflin, Boston, 1921.

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

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, et al. Natural emergent misalignment from reward hacking in production RL. https://assets.anthropic.com/m/74342f2c96095771/original/Natural-emergent-misalignment-from-reward-hacking-paper.pdf, November 2025.

Antonio Martini, Terese Besker, and Jan Bosch. Technical debt tracking: Current state of practice: A survey and multiple case study in 15 large organizations. *Science of Computer Programming*, 163:42–61, 2018. doi: 10.1016/j.scico.2018.03.007.

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, December 2024. URL https://arxiv.org/abs/2412.04984.

Munich Re. Cyber insurance: Risks and trends 2025. https://www.munichre.com/en/insights/cyber/cyber-insurance-risks-and-trends-2025.html, 2025. Accessed 2026-02-12.

Milad Nasr, Nicholas Carlini, Chawin Sitawarin, Sander V. Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, Shuang Song, Harsh Chaudhari, Ilia Shumailov, Abhradeep Thakurta, Kai Yuanqing Xiao, Andreas Terzis, and Florian Tramèr. The attacker moves second: Stronger adaptive attacks bypass defenses against LLM jailbreaks and prompt injections. *arXiv preprint arXiv:2510.09023*, 2025. URL https://arxiv.org/abs/2510.09023.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025. URL https://arxiv.org/abs/2505.23836.

OpenAI. Our approach to alignment research. https://openai.com/index/our-approach-to-alignment-research/, August 2022. Accessed 2026-02-12.

OpenAI. Introducing superalignment. https://openai.com/index/introducing-superalignment/, July 2023. Accessed 2026-02-12.

OpenAI. Continuously hardening ChatGPT Atlas against prompt injection attacks. https://openai.com/index/hardening-atlas-against-prompt-injection/, December 2025a. Accessed 2026-02-12.

OpenAI. Preparedness framework. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf, April 2025b. Version 2.0, last updated 15 April 2025.

PromptArmor. Data exfiltration from Slack AI via indirect prompt injection. PromptArmor Blog, August 2024. URL https://promptarmor.substack.com/p/data-exfiltration-from-slack-ai-via.

Johann Rehberger. Microsoft Copilot: From prompt injection to exfiltration of personal information. Embrace The Red (blog), 2024. URL https://embracethered.com/blog/posts/2024/m365-copilot-prompt-injection-tool-invocation-and-data-exfil-using-ascii-smuggling Demonstrates chaining prompt injection, automatic tool invocation, ASCII smuggling, and hyperlink rendering to exfiltrate data from Microsoft 365 Copilot. Disclosed January 2024; patched July 2024.

Kevin Roose. A conversation with Bing's chatbot left me deeply unsettled. *The New York Times*, February 2023. URL https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015. URL https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html.

Steven Shavell. The judgment proof problem. *International Review of Law and Economics*, 6(1): 45–58, 1986. doi: 10.1016/0144-8188(86)90038-4.

Joseph Spracklen, Raveen Wijewickrama, A H M Nazmus Sakib, Anindya Maiti, Bimal Viswanath, and Murtuza Jadliwala. We have a package for you! A comprehensive analysis of package hallucinations by code generating LLMs. In *USENIX Security Symposium*, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/spracklen.

Jacob Steinhardt. AI forecasting: One year in, July 2022. URL https://bounded-regret.ghost.io/ai-forecasting-one-year-in/. Blog post, Bounded Regret.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=7Qa2SpjxIS.

Simon Willison. The lethal trifecta for AI agents. Simon Willison's Weblog, June 2025. URL https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/. Also available at https://simonw.substack.com/p/the-lethal-trifecta-for-ai-agents.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*, October 2023. URL https://arxiv.org/abs/2310.02446. NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR) 2023.

A Statement on LLM Usage

In accordance with ICLR's policy on the use of large language models (LLMs) during paper preparation, we disclose the extent and nature of LLM involvement in this work: In preparing this work, the authors used Claude (Anthropic) and Chat Generative Pre-Trained Transformer (ChatGPT, OpenAI) to assist with brainstorming, literature research, drafting, and iterative refinement. The conceptual framework — AI safety debt as an extension of technical debt — was developed by the authors. LLM assistance included identifying relevant examples from the research literature, generating initial drafts of sections for use as starting-points, and suggesting structural alternatives during revision. LLM-generated content was critically evaluated; suggestions were frequently rejected and substantially modified to reflect the authors' views. The authors reviewed all content and take full responsibility for the final publication.