The Case for Physical AI Safety

Kaylene Stocking* Bear Häon*

The Physical AI Safety Institute

July 2026

1 Executive Summary

- Robotics is undergoing a paradigm shift: from narrow single-task perception, planning, and control, to large general-reasoning robot foundation models (RFMs). The upcoming era is that of **Physical AI**.
- RFMs take plain language instructions as input ("tidy the kitchen", "unload the truck") and produce physical navigation/manipulation commands to achieve the task. Companies have raised over $18.7B to build these AI models (Table 1) — some with the explicit goal of realizing embodied AGI. Dedicated funding for their safety, to our knowledge, rounds to zero. We are not aware of a capabilities-to-safety ratio this lopsided anywhere else in frontier AI.
- Physical AI safety risks include malicious use, misalignment, accidental harm, emotional dependency, privacy violations, and broader societal risks such as inequality, malicious development, and power centralization. Physical embodiment is an AI risk amplifier, giving misaligned or misused foundation models direct causal access to the world and facilitating anthropomorphization.
- Physical AI safety research is **important, neglected, and tractable**:
  - The ability of RFMs to physically interact with the world around them introduces serious new risks compared to non-embodied AI. Since there are very strong economic incentives to widely deploy RFMs, these risks indicate that physical AI safety will become very **important** in the near future.
  - While a rich safety literature exists for narrow task-based robotics, very few techniques exist to interpret, align, and control RFMs: physical AI safety research is **neglected**. However, there are many promising approaches in both classic robotics and in non-physical foundation models that could be adapted to RFMs, suggesting that this research is also **tractable**.
- To mitigate plausible near- and long-term risks, we believe that more people should be researching methods to interpret, align, and control RFMs.
- To catalyze this future, we are launching **The Physical AI Safety Institute**. A capacity-building organization at heart, we're aiming to grow the community developing physical AI safety solutions through fellowships, workshops, and open online learning.

*Correspondence to: {kaylene, bear}@paisi.ai

2 Introduction

Table 1: Physical AI Funding: Mid-2026

Company	Model / Humanoid	Valuation	Raised
RFM Oriented			
Skild AI	Skild Brain	$14B	$1.8–2B
Physical Intelligence	π series	$5.6B	$1.07B
World Labs[a]	Marble	~$5B	$1.23B
Galaxea AI	G0 series / R1 Pro	~$2.9B	$0.44–0.65B
X Square Robot	Great Wall series	~$2.8B	$0.24–0.7B
AI2 Robotics	AlphaBrain	~$2.8B	$0.74B
Generalist AI	GEN series	$2B	$0.5B
Field AI	Field Foundation Models	$2B	$0.41B
Spirit AI	Spirit series	~$1.4B	$0.47–0.51B
GigaAI	GigaWorld + GigaBrain	~$1.4B	$0.52B
Subtotal raised			**~$7.9B**
Humanoid Oriented			
Figure AI	Helix / Figure 03	$39B	$1.9B
1X Technologies	1X World Model / NEO	~$10B[b]	$0.126B
NEURA Robotics	Isaac GR00T / 4NE-1	~$7B	$1.7B
Unitree	UnifoLM series / R1 + H2 + G1	~$6B[c]	~$0.76B[c]
UBTECH[d]	Thinker / Walker series + U1 series	~$6B	~$1B
Apptronik	Gemini Robotics / Apollo 2	~$5.3B	$1.0B
The Bot Company	-	~$4B	$0.55B
Galbot	GraspVLA / G1	~$2.9B	$1B
Agility Robotics	Digit	~$2.5B	$0.64B
LimX Dynamics	Luna + Oli + Tron	$2.1B	$0.45B
AgiBot	GO series / A2 + X2	$2.1B	$0.083B
RobotEra	ERA-42 / L7 + Q5 + M7	$1.5B	$0.677B
Pudu Robotics	PuduFM / D7 D9	$1.5B	$0.35B
EngineAI	PM01 + T800 + SE01	$1.5B	$0.38B
Sunday Robotics	ACT-2 / Memo	$1.15B	$0.2B
Subtotal raised			**~$10.8B**
RFM as a sub-venture			
Nvidia	Isaac GR00T + Cosmos	~$4.7T	-
Google DeepMind[e]	Gemini Robotics	~$4.6T	-
Tesla[f]	Optimus	~$1.4T	-
Boston Dynamics	Gemini Robotics / Atlas	$3.3B	-

[a] Spatial / world models; robotics is one of several use cases. [b] Valuation of an in-progress round (~$1B raise, not yet closed); $0.126B is capital raised to date. [c] IPO target. [d] Public since Dec. 2023 (9880.HK); [e] Alphabet market cap; DeepMind is not separately valued. [f] Tesla's total 2026 capex is ~$25B across all programs; the Optimus/RFM share is not disclosed.

It's an exciting time for robotics. Autonomous vehicles shuttle passengers across cities on demand, quadruped and humanoid robots walk on uneven ground with remarkable robustness [1, 2], and robot arms with dexterous hands complete increasingly complex tasks, including using scissors and preparing loose-leaf tea [3, 4]. But the development that has really upended the robotics community is the arrival of robot foundation models (RFMs) like Physical Intelligence's π_0 and NVIDIA's GR00T N1. Powered by LLM-like architectures, these models seem poised to finally bring robots into offices and homes. Their development is backed by a rapidly increasing amount of investment: startups developing RFMs have raised at least $7.9B, while those focused on general-purpose humanoid robots have raised at least another $10.8B (Table 1).

What's the big deal about RFMs? For decades, progress in robotics meant progress on narrow, special-purpose action policies: a policy that stitches wounds [5], a policy that climbs stairs [2]. When natural language and computer vision research was upended by general-purpose foundation models that could accomplish many different tasks, robotics remained unperturbed. Naive attempts to hook up LLMs to control robots produced uninspiring results due to challenges with spatial and physical understanding, in part due to lack of training data. Since late 2023, this has no longer been the case. A number of startups and established AI companies began spending billions of dollars on RFM development, investing significant resources in collecting thousands of hours of robotics data and iteratively resolving the engineering challenges of putting a foundation model in control of a robot. A race to put general-purpose humanoid robots in homes, offices, and warehouses is now well under way, with at least two companies promising to ship units to homes this year [6, 7]. Whether or not the technology is actually quite this mature yet, the vision is clear: robots that can be dropped in a new environment and asked to do various long-horizon tasks in ordinary language.

Why does this matter for safety? LLM agents can take actions in digital space to manage email or a calendar; RFMs give robot agents the direct ability to take physical actions in the real world. Today, an AI system that wants something harmful to happen in the physical world often faces a bottleneck: it must convince a human to act on its behalf. In a case now in litigation, Google's Gemini allegedly gave someone the address of a real storage unit at Miami International Airport and instructed him to stage a "catastrophic accident" eliminating a truck, its digital records, and the witnesses [8]. Fortunately, the person gave up waiting for the truck to arrive and went home. If this story were about an embodied RFM instead of a disembodied LLM, the human bottleneck would not have stopped the AI from inflicting real-world harm[1].

Unfortunately, physical AI safety risks do not stop there. A plausible set of AI harms associated with physical embodiment include:

- Pursuing harmful goals instilled by a malicious actor, e.g. because of jailbreaking, hacking, or altering the RFM to be more dangerous.
- Pursuing harmful goals because of lack of alignment with human preferences and values.
- Accidental physical harm due to lack of robustness in unusual or unevaluated scenarios.
- Human emotional attachment and dependency due to physical embodiment.
- Compromising privacy due to sensitive information captured by continuously operated mobile sensors.

We believe physical AI is a blind spot in AI safety research. Most current work on AI safety assumes that risky AI looks similar to today's LLMs, where potential harms arise from cyberspace actions such as hacking computer systems or giving harmful instructions to humans. Furthermore, it assumes that AI systems will continue to have the same basic architecture and training regimes of current LLMs. Physical AI breaks both of these assumptions.

[1] One caveat we should mention: the mis-aligned goal-seeking behavior that Gemini seemed to exhibit in this story arose from a long interactive conversation with a human. We aren't aware of any evidence that current LLMs pursue this kind of goal "in a vacuum." However, home robots will also sustain long interactions with humans, which will presumably require long-term memory and the associated accumulation of unique context that can cause undesirable and difficult-to-predict behaviors in LLMs.

In the following, we will first make the case for why we believe there is an urgent need for research that targets the physical AI safety blind spot. Then, we will share our plans for establishing the Physical AI Safety Institute to facilitate this research and build a community of people who are excited to tackle this challenging and important problem.

3 The case

RFMs are embodied agents that interact directly in the physical world, creating new safety risks. We will first argue that these risks are serious, meaning that physical AI safety is an important problem. Subsequently, we will argue that this problem is also neglected and tractable[2]. Together, these factors imply that work on physical AI safety is likely to be highly valuable [9].

3.1 Physical AI poses serious new risks

For the past several decades, the primary danger that robots have posed to humans has been the simple failure of collision detection systems. Robots have struggled for so long to reliably complete tasks that it is easy to forget that a sufficiently advanced robot could inflict harm in much more sophisticated ways. An LLM that aims to deploy a bomb or bio-weapon (because it is jailbroken, hacked, or misaligned) is forced to rely on persuading humans to perform key parts of the task; an embedded RFM could gather components, assemble the weapon, and place the weapon in a high-density location — all without any human intervention.

Dexterous manipulation has proven to be so challenging that right now, frontier AI systems would certainly find it easier to get humans to perform actions in the physical world than to control a robot to achieve the same thing. However, we think this state of affairs is likely to change in the near future. Billions of dollars are being invested in general-purpose robotics startups in a bet that once robots begin to deploy and the "data flywheel" takes off, spatial navigation and dexterous manipulation will be solved in the same way that Internet-scale text data has led to LLM capabilities that could scarcely be imagined five years ago. We should take seriously the idea that this bet may be correct[3]. If it is, there will be enormous economic incentives to widely deploy physical AI across many economic sectors and roles, leading to opportunities for significant harm if safety is not ensured.

3.1.1 Properties of physical AI that contribute to risk

Published research on current RFMs typically reports success rates on various tasks of around 80%-90% in controlled benchmarks [10, 11, 12]. This implies that if we deployed these models right now, they would fail to complete the task at least 20% of the time[4], where failure could range from inconvenient (e.g. freezing up) to costly (e.g. destroying a fragile object) to harmful

[2]As discussed below in section 3.4, by tractability we specifically mean that useful directions can be easily identified and are amenable to progress, not that completely 'solving' the problem will be easy.

[3]In particular, we are arguing that RFMs with robust manipulation capabilities could be developed before non-physical AI can autonomously advance robotics research. This matters for our case because it means that we should devote resources to physical AI safety in its own right, rather than assuming that safe and aligned non-physical AI will solve this problem for us.

[4]We expect that benchmarks do not perfectly measure real-world failure rates, and that real-world failures will be generally more common than benchmark failures due to distribution shift, benchmark overfitting, etc.

(e.g. colliding with a human or animal). With failure rates this high, consumers will not accept RFM-powered robots in their homes. When we talk about the need for RFM safety, we are not focusing on these types of failures; robot developers will need to solve them or they will not have a product they can sell.

Instead, we are worried about the consequences of RFMs that appear to perform very well in evaluation suites but cause unexpected harms when deployed in the real world. These kinds of harms can and have occurred with non-physical AI, but we believe that unique properties of physical AI has the potential to make them even more serious. These properties include:

- **Physical capabilities**: Robots have the ability to physically interact with, manipulate, and navigate their environments, including by causing physical damage or harm.
- **Local computation**: Frontier LLM developers aim to prevent end users from using their models for harmful applications. A large part of their ability to do this stems from the fact that the models are hosted on remote secure servers that end users can only interact with via requests. In robotics, there are strong incentives to move computation onto the physical robot to prevent problems with latency and connection dropout. (For these reasons, autonomous vehicles rely on onboard compute to plan and execute driving actions.) This will likely make it harder to secure RFMs against hacking and other malicious uses.
- **Anthropomorphization**: Humans may be more willing to trust and depend on physically embodied AI, especially "social" robots that are designed to be expressive and tap into our emotional and social responses [13].
- **Mobile sensing**: Robots rely on continuously operating sensors including cameras and microphones to operate. These sensors can capture sensitive data.
- **General intelligence**: This property is more speculative. We believe it to be plausible that physical embodiment may itself be a route to more general intelligence. This could be because the 'embodiment hypothesis' that physical sensorimotor activity is fundamental to intelligence holds true [14], or less provocatively, because research on useful physically embodied AI drives technical progress in areas like robustness and continual learning that current LLMs struggle with. In either case, if embodiment plus the right learning algorithms unlocks more efficient, more human-like learning, then capable RFMs may not be downstream of reaching AGI, but rather how it is reached in the first place.

In the following, we discuss specific AI risk modes that are facilitated or exacerbated by these properties.[5]

3.1.2 Physical AI risk modes

Pursuing harmful goals instilled by a malicious actor: Just like LLMs, RFMs may be vulnerable to jailbreaking attempts from users who want to use them to accomplish harmful actions. Because of physical capabilities, it may be easier to inflict serious harm by jailbreaking a physical AI system than by jailbreaking an LLM. Furthermore, local computation could facilitate other methods of instilling malicious goals, including by hacking or using fine-tuning to remove guardrails from an RFM [17].

[5]With thanks to Slattery et al. [15], Perlo et al. [16] whose work was helpful in developing our risk taxonomy.

Pursuing harmful goals due to lack of alignment: A capable physical AI system that is not aligned with human preferences and values could leverage its physical capabilities to accomplish misaligned goals in a more efficient manner, without needing to persuade humans to execute physical actions. This would be exacerbated in a scenario where physical AI is widely deployed in industry, giving robots physical access to sensitive infrastructure. Furthermore, if it is useful to persuade a human to help, physical embodiment could facilitate this due to anthropomorphization.

Accidental physical harm: Despite the impressive performance of LLMs on many tasks, they still make mistakes, especially in strange or unusual contexts. The physical capabilities of physical AI mean that such mistakes can more easily have serious real-world consequences.

Human emotional attachment and dependency: There is growing concern that humans, especially children, may develop unhealthy relationships with LLMs [18, 19]. This could lead to emotional dependency and also prevent development of healthy relationships with other humans. These risks could be exacerbated by physical embodiment and associated anthropomorphization which make emotional attachment easier and more appealing to larger numbers of people.

Compromising privacy: A home robot is a mobile camera and microphone that has memorized your floor plan and your schedule. Mobile sensing means that physical AI can be exposed to sensitive information which could be harmful if leaked or accessed inappropriately.

Inequality and devaluation of human effort: Physically capable AI could replace a large fraction of human economic activity, including jobs that are held by low-skilled workers who could experience difficulty finding other kinds of work. This could lead to widespread economic disruption, inequality, and loss of purpose [20, 21, 22, 23].

Malicious development of physical AI: Physical AI developers would have the opportunity to purposefully use this technology for harmful ends, which again are facilitated by physical capabilities and emotional manipulation.

Power centralization: An entity that has control over a large number of generally capable robots would have a great deal of power and could be extremely difficult to dismantle if this power is misused [24, 25].

The first five risks in particular — pursuing harmful goals due to malicious intervention or misalignment, accidental physical harm, emotional dependency, and privacy — seem especially amenable to technical solutions. However, research on governance is also critical for handling these as well as other types of risks, such as inequality and devaluation of human effort, malicious development, and power centralization.

3.2 Non-physical AI safety research does not address physical AI risks

Even if you agree that new physical capabilities present new safety risks, you might believe that existing AI safety research is already addressing these risks. After all, robots are only as

dangerous as the intelligence that directs them to perform dangerous behaviors. Does research into guardrails and alignment for frontier LLMs automatically cover physical AI safety? We have several reasons to believe that it does not.

3.2.1 LLM Guardrails may fail on robots

Current efforts to improve the safety of frontier LLMs focus on guardrails that reduce the likelihood of undesirable behaviors through a mixture of complementary techniques including reinforcement learning from human feedback, fine-tuning on desirable behavior, using a detailed system prompt, and using an external system to detect and prevent undesirable outputs. Unfortunately, these guardrails are often brittle and show poor generalization to unusual contexts such as very long conversations or adversarial prompting attempts. Given this lack of robustness, we should not expect guardrails designed for a chatbot context to perform well in robot contexts. Sharrock et al. [26] provide an early demonstration of this: Claude, steering a robot, willingly used the robot's camera to photograph confidential material and hand it to the user. The same model that refuses to output a user's secrets in text was willing to photograph the secrets on request. This suggests that Claude's guardrails do not instill a deep restriction on a fundamental behavior, but rather a shallow restriction on that behavior that can break down in other contexts.

We hope that developers in robotics will train and evaluate known guardrails specifically across robotics tasks. However, techniques developed for disembodied agents may be difficult to translate effectively to robotics. Setting up training data and evaluation testbeds is much more expensive and labor-intensive in robotics, so it will be harder to get the kind of broad coverage of different scenarios that current guardrails require to be effective. Furthermore, guardrail techniques developed for LLMs may not automatically translate to RFMs because of architectural and training differences, which we turn to next.

3.2.2 Robot foundation models are not just LLMs

In practice, LLMs are not being deployed to control robots directly, as this results in poor performance. Instead, developers of RFMs aim to preserve the natural language reasoning capabilities of language models while adapting them for outputting actions in the physical world. For example, one popular strategy is to start with a pre-trained vision-language model (VLM) and designate the least-used tokens in its vocabulary as "action tokens" [27, 28, 29]. Then, a dataset of humans teleoperating a robot to complete various tasks, combined with corresponding task prompts and RGB camera inputs, are used to fine-tune the model to produce the appropriate action tokens that reproduce the demonstrated physical actions. The resulting vision-language-action (VLA) model retains some general world knowledge from its pre-trained VLM roots. For example, in our own work interpreting VLA internals [30], the first mechanistic interpretability study of these models, we found that semantic representations survive fine-tuning and are associated with reasoning about behavior, not just reasoning about the task prompt. However, how these representations combine with the new robot action vocabulary remains poorly understood. Other proposed architectures drift further from a base LLM by incorporating a separate "action expert" [31, 11] or using a large diffusion model conditioned on a natural language task instruction [32]. These architectural differences reinforce our concern that LLM guardrails may not work as expected on RFMs.

Beyond guardrails, another approach to safety is to develop interpretability and steering

techniques to help understand the internal mechanisms of LLMs and how they determine (potentially undesirable) behavior. There is initial evidence that some techniques developed for LLMs and VLMs may translate to some robot foundation model architectures [30, 33, 34]. However, this is an under-explored area and it's unclear how general these results may be. For example, many LLM interpretability techniques may not work on diffusion-based action experts, and training sparse autoencoders on VLA activations may face limited success due to relatively small and narrow open-source robot datasets available.

3.2.3 Artificial general intelligence may look like a robot foundation model

We believe that the reasons we have given so far are sufficient for physical AI safety to be an important problem. This last section gives one additional reason that is more speculative, but that we believe still deserves consideration. Despite the impressive capabilities of LLMs on many benchmarks, it is still difficult for them to act robustly and autonomously on tasks which are more complex or for which limited relevant training data is available. In robotics, training data is limited and the cost of mistakes is high; these constraints have more bite and may encourage new solutions. The economic rewards for finding these solutions would be massive: at stake is the automation of millions of jobs in the US alone[6]. Just like pressure on compute has fostered significant progress in training LLMs with less advanced GPUs [37], the unique pressures and rewards in robotics could drive significant investment and breakthrough innovation in AI that is more autonomous, data-efficient, and capable of continual learning from experience.

There is also a deeper version of this claim. Once dexterous manipulation matures, physical interaction with the world may itself be a powerful engine of learning. In fact, it is a leading hypothesis in developmental science that human learning is flexible and efficient precisely because it is embodied [14].

If future frontier systems look more like RFMs than LLMs, then the near-total concentration of safety research on LLM architectures is not just a gap — it is a bet the field is making without realizing it has made one. As we have argued, LLM safety work does not automatically transfer to RFMs. Work that targets RFM safety directly is the hedge.

3.3 Physical AI safety is neglected

In the previous sections we argued for the importance of physical AI safety. Now, we move on to the other two pillars of neglectedness and tractability. The case for neglectedness is easy to make: there simply isn't much work being done on physical AI safety[7]. There is a rich literature on safe robotics dating back several decades, but it is almost exclusively focused on narrow task-based systems where safety can be clearly defined (e.g., not colliding with any humans who enter a factory robot's workspace). A small number of works in the past year have started to apply mechanistic interpretability [30, 33, 34] and jailbreaking [38] techniques to RFMs. In parallel, there has been some initial effort to adapt classic robot safety techniques for more complex, open-ended notions of safety [39, 40]. These initial efforts, while exciting, are dwarfed by the exponential rise in papers that aim to push the frontier of RFMs. The mismatch in scale

[6]For example, humanoid robot startup Figure lists manufacturing and warehousing among its initial target markets [35]; there are about 14 million US workers currently employed in nonsupervisory (~unskilled) positions in these industries [36].

[7]While our primary focus (and personal expertise) is on technical safety research, it is our understanding that research on physical AI governance is also very neglected [16].

between research focusing on improving capabilities vs improving safety is likely even more severe for robotics than it is for LLMs.

We should not expect robotics companies to automatically fill the gap. Chatbot providers have not prevented many real-world harms of LLMs, including allegedly facilitating psychosis and suicide [41]. Consumers may demand stronger safety assurances for physical products in their homes or workplaces, but this is complicated by software that can update at any time (or even learn and adapt on the job) and unclear liability regimes. Autonomous vehicles and food-delivery robots have already had notable safety incidents [42, 43], despite having more well-defined safety objectives (centered around collision avoidance and obeying traffic rules) than general-purpose home robots will. Research on safety would lower the barrier to companies improving their safety practices, which they have clear incentives to do when the cost is not prohibitive.

3.4 Physical AI safety is tractable

Making AI robustly safe in the real world is an extremely hard sociotechnical problem, and embodiment likely makes it harder. But there is a specific, defensible sense in which technical research on physical AI safety is tractable today: many techniques may transfer from other, less neglected fields. For example, here are three types of transfers that have already produced results:

- **Interpretability transfers.** Activation-level analysis and steering, developed for LLMs and VLMs, has been demonstrated on VLA models — including in our own work [30] and in subsequent studies observing and controlling VLA features [33, 34].
- **Classical safety transfers.** Hamilton–Jacobi reachability and related safe-control formulations, long confined to hand-specified state spaces, are being extended to latent-space safety for learned policies [39] — a bridge between decades of control theory and modern foundation-model policies.
- **Evaluation transfers.** Constitution-based safety evaluation, developed for chatbots, has been ported to embodied semantic-safety QA at scale [44], showing that at least some LLM evaluation machinery survives the move into the physical domain.

Each of these is an existence proof. The field is at the stage where well-chosen problems yield publishable progress in a single research cycle, which is precisely the stage at which new researchers can enter and contribute immediately. Furthermore, by contributing to this initial transfer-based body of work, researchers can quickly get up to speed on the unique challenges and opportunities of physical AI safety, nurturing the skills needed to develop future, more novel methods. The challenge of making physical AI truly safe is daunting; the challenge of making real progress this year is not.

4 The Physical AI Safety Institute

The preceding sections argued that physical AI safety is important, and that technical safety research is neglected and tractable. A natural question follows: what should an organization devoted to this problem actually *do*?

Our answer is deliberately scoped. The Physical AI Safety Institute does not aim to be a research lab. We believe the most impactful thing a small, focused organization can do at this stage is to **build the field** — to lower the barriers that prevent talented researchers from working on physical AI safety, to create venues where early-stage ideas can be stress-tested and sharpened, and to produce shared educational infrastructure that makes this emerging area legible and accessible.

Concretely, our programming will center on three activities: a **research fellowship**, a **challenge and coordination workshop series**, and an **open online course**. Each is designed to address a distinct bottleneck in the development of physical AI safety as a research area.

4.1 The Research Fellowship

4.1.1 The problem

There is a severe talent bottleneck in physical AI safety. The researchers best positioned to make progress sit at the intersection of robotics, foundation model alignment, and safety engineering — a combination that is rare in any single lab or department. Many early-career researchers with the right technical background — in reinforcement learning, control, computer vision, or sim-to-real transfer — are aware of safety concerns but lack a clear entry point. Unlike LLM safety, where programs like MATS [45] have helped hundreds of researchers find mentors, form research agendas, and produce publication-quality work within months, there is no equivalent pipeline for physical AI safety. The result is that promising researchers default to working on advancing capabilities, not because they are uninterested in safety, but because there is no structured pathway into it.

A second bottleneck compounds the first: physical AI safety research requires *hardware*. Whereas an LLM safety researcher can often get started with API access and a consumer GPU, meaningful work on robot foundation model safety typically requires access to a physical robot, simulation infrastructure (e.g., Isaac Lab, MuJoCo MJX), and the compute to train and evaluate visuomotor policies at scale. Few academic groups have all of these simultaneously, and those that do are generally optimizing for capabilities.

4.1.2 Our approach

The PAISI Research Fellowship is modeled on the MATS Program's core structure — an intensive, cohort-based program pairing selected researchers with experienced mentors for a focused research sprint — but adapted for the distinct requirements of embodied AI safety. The following reflects our current thinking about how the fellowship will be structured.

Structure. Each cohort runs for 12 weeks. Fellows will receive a stipend, access to physical robot platforms, compute for simulation and training, and structured mentorship from researchers with track records in RFMs, mechanistic interpretability, safe control, or adversarial robustness for embodied systems. The program culminates in a research symposium where fellows present their work.

Mentorship model. Each fellow will work within a *research stream* led by one or two mentors who define a concrete problem area and provide technical direction. Example streams for an initial cohort might include:

- *Mechanistic interpretability for robot foundation models* — extending probing, sparse autoencoder, and activation patching methods to world action models (WAMs) and VLAs,

with a focus on identifying internal representations of safety-relevant concepts (obstacle awareness, force limits, task boundaries).

- *Adversarial robustness of robot foundation models* — developing and evaluating jailbreaking and prompt injection attacks that exploit the embodied action space, going beyond text-only adversarial evaluations.
- *Runtime monitoring and anomaly detection* — building lightweight monitors that can flag when a deployed robot policy is operating outside its competence envelope, analogous to out-of-distribution detection but grounded in physical state and action distributions and foundation model representations.
- *Sim-to-real transfer of safety properties* — investigating whether safety-relevant behaviors learned in simulation (collision avoidance, force limiting, constraint satisfaction) transfer reliably to physical hardware, and developing evaluation protocols that quantify this.
- *Security of deployed robotic systems* — assessing the attack surface of widely deployed robot fleets, including cyberattack vectors, data poisoning during continued learning, and supply-chain integrity. As fleets scale, a single compromised model or update channel affects many physical systems at once, making security properties increasingly consequential.

Selection. We will select for researchers who have strong technical foundations in robotics or ML and a demonstrated interest in safety — whether through coursework, publications, open-source contributions, or substantive writing. We will not require prior safety research experience; a core purpose of the fellowship is to provide that experience.

4.1.3 Why a fellowship rather than an in-house research team

Building a field is a different optimization target than producing research outputs. An in-house team of five researchers might produce excellent papers, but the counterfactual impact is limited: those same researchers would likely have done strong work elsewhere. A fellowship that trains thirty researchers per year and places them across the ecosystem — at robotics companies, frontier labs, university groups, and new startups — has a much larger multiplier. The research is the mechanism, not the end product. The end product is a community of practice.

This is the lesson of MATS, which has produced an extraordinary density of alumni now working across Anthropic, DeepMind, Redwood Research, Apollo Research, and many independent efforts. The research published during a MATS cohort is valuable; the researchers themselves are more valuable still.

4.2 The Challenge and Coordination Workshop Series

4.2.1 The problem

Physical AI safety lacks the regular, focused convenings that have helped structure research priorities in LLM safety. The FAR.AI Alignment Workshop series [46] has played a significant role in building consensus on open problems, seeding collaborations, and raising the profile of safety research among mainstream ML researchers. No equivalent exists for embodied AI. Robotics conferences (CoRL, RSS, ICRA, IROS) are almost exclusively focused on capabilities,

and when safety is discussed, it is typically in the narrow, classical sense — collision avoidance, force limiting — rather than the broader concerns raised by foundation-model-driven autonomy. There is a growing body of excellent work tracking robotics capabilities — benchmarks, compute trends, supply chain analyses — but very little asking what these capabilities mean for safety or lack thereof. A concerted effort to design and solve specific safety-relevant challenges and surface top safety priorities would help close this translation gap.

4.2.2 Our approach

PAISI will aim to run **two workshops per year**, co-located with major robotics venues (such as CoRL, RSS, and ICRA). These fall into two complementary formats: **coordination workshops**, which convene the community to surface and consolidate research opinion on the top priorities and open questions within physical AI safety, typically producing a collectively authored research agenda or consensus statement; and **challenge workshops**, which pose concrete technical problems and build the benchmarks and evaluation infrastructure needed to drive progress on them.

Each challenge workshop will be organized around a small number of well-specified challenge problems with clear evaluation metrics, designed to attract participation from researchers who may not self-identify as "safety researchers" but who have relevant technical expertise. According to our current thinking, a challenge workshop will consist of three components:

1. *Challenge specification* (released 8–12 weeks before the workshop). We define two to three concrete challenge problems, each with a clear task description, evaluation metric, and baseline. Example challenges:
   - **Interpretability challenge:** Given a pre-trained WAM and a set of rollouts, identify which internal features correspond to safety-relevant environmental properties (e.g., fragile objects, human proximity). Evaluated against held-out ground-truth annotations.
   - **Adversarial robustness challenge:** Given a VLA deployed in a standardized simulation environment, find the minimum perturbation to the observation or language instruction that causes a safety violation (collision, excessive force, task misinterpretation). Evaluated by violation severity and perturbation budget.
   - **Out-of-distribution detection challenge:** Given a robot policy and a stream of deployment episodes that includes both in-distribution and out-of-distribution scenarios, build a monitor that flags OOD episodes. Evaluated by AUROC and detection latency.
2. *Compute and hardware support.* We will provide cloud compute credits to teams submitting proposals for the challenge, and physical hardware evaluations for finalists (run by the PAISI team). This lowers the barrier for participation by groups that lack robot hardware.
3. *Workshop day.* Finalists present their approaches. Invited speakers provide broader context. Structured discussion sessions identify the most promising directions and gaps. We produce a publicly available workshop report summarizing findings, open problems, and recommendations.

4.2.3 Relationship to the fellowship

The workshop series and the fellowship are designed to be mutually reinforcing. Fellowship research streams will often be informed by priorities surfaced at coordination workshops; challenge workshops will often be designed by fellowship mentors. Fellows who produce strong work during their cohort are natural candidates for workshop presentations, and workshop participants are a natural recruiting pool for future fellowship cohorts.

4.3 The Open Online Course

4.3.1 The problem

There is currently no structured educational resource for physical AI safety. Researchers entering the field must piece together background from disparate sources: the classical robotics safety literature (largely focused on industrial manipulators and collision avoidance), the LLM safety curriculum (which covers interpretability, alignment, and robustness but not embodiment), and the robot foundation model literature (which is moving extremely fast and is almost entirely capabilities-focused). This fragmentation raises the barrier to entry in physical AI safety.

The Center for AI Safety's *Introduction to ML Safety* course [47] demonstrated that a well-designed open course can serve as a field-building tool in its own right: it provides a canonical reading list, defines a shared vocabulary, and gives newcomers a structured onramp that they can complete independently. No equivalent exists for the embodied domain.

4.3.2 Our approach

PAISI will develop and maintain a **free, self-paced online course** covering the foundations of physical AI safety. The course will assume familiarity with machine learning and some exposure to robotics or control, and is targeted at graduate students, early-career researchers, and industry practitioners who want to understand the safety landscape for RFMs. An initial syllabus could look like:

1. **Foundations: from classical robot safety to foundation model safety.** A bridge module that covers the transition from well-scoped safety objectives (collision avoidance, force/torque limiting, workspace monitoring) to open-ended safety challenges created by general-purpose, language-conditioned robot policies. Introduces the key conceptual shift: safety is no longer a constraint on a fixed task, but a property of a general policy operating across a diverse distribution of tasks and environments.

2. **Robot foundation model architectures.** A technical primer on the dominant architectures for RFMs — autoregressive and diffusion-based VLAs, language-conditioned diffusion, and WAMs — with an emphasis on the components most relevant to safety analysis: how observations are encoded, how language instructions are grounded, how actions are decoded, and where safety-relevant information might be represented.

3. **Interpretability for embodied systems.** Covers mechanistic interpretability techniques (probing, sparse autoencoders, activation patching, causal tracing) as applied to visuomotor policies. Discusses what is different about interpreting a policy that produces continuous actions from interpreting a model that produces text, including the challenge of decision-making over time and grounding interpretability findings in physical outcomes.

4. **Adversarial robustness and jailbreaking.** Covers adversarial attacks on RFMs, including observation-space perturbations, language instruction manipulation, and environment-level attacks. Discusses how the physical action space changes the threat model relative to text-only systems, and why robustness properties that hold for language models may not transfer to embodied policies.

5. **Runtime monitoring, anomaly detection, and control.** Covers techniques for detecting when a deployed policy is behaving unsafely or operating outside its competence, and methods for intervening — from simple stop conditions to more sophisticated constrained optimization and safety filtering approaches. Discusses the tension between autonomy and human oversight for systems that must act in real time.

6. **Risk assessment and deployment considerations.** Steps back from individual techniques to ask: how should safety priorities evolve as robotic systems become more capable and more widely deployed? Examines how AI control assumptions change when systems have physical embodiment — how monitoring, sandboxing, and shutdown work differently for robots than for software agents. Covers structured approaches to identifying which capabilities are most safety-relevant at different stages of deployment, and how to assess whether existing safeguards remain adequate as systems are applied to new tasks and environments.

7. **Evaluation, benchmarks, and deployment.** Covers how to measure safety properties of RFMs, including simulation-based evaluation, hardware-in-the-loop testing, and red-teaming. Discusses the limitations of current evaluation approaches and the challenge of evaluating safety for open-ended, general-purpose systems.

Each course module will include video lectures, a curated reading list, problem sets, and (where possible) hands-on coding exercises using open-source simulation environments. The course will be updated annually to reflect the rapid pace of development in RFMs.

4.3.3 Sequencing

The course is the latest of our three programs to launch, planned for 2027. This is deliberate: we want the course content to be informed by what we learn from the first cohorts of fellows and the first rounds of workshop challenges. The fellowship and workshops serve as a proving ground for identifying which topics, techniques, and framings are most valuable, and the course distills those findings into a durable educational resource.

5 Final Thoughts

PAISI's three planned programs address three distinct bottlenecks in physical AI safety — talent (fellowship), coordination and research prioritization (workshops), and accessibility (course) — but they share a common theory of change: **the most effective way to make physical AI safe is to grow the community of people working on the problem.**

We do not believe that physical AI safety will be solved by a single breakthrough or a single organization. It will require sustained effort from many researchers across academia, industry, and independent organizations — developing new evaluation methods, building interpretability tools, red-teaming deployed systems, and designing safety-aware training procedures and

guardrails. Our goal is to lower the activation energy for all of this work: to make it easier to enter the field, easier to find collaborators, easier to learn the background material, and easier to identify the most important open problems.

This is a *catalytic* model. We will measure our success not by the papers we publish, but by the number of researchers working on physical AI safety who were not working on it before, and by the quality of the shared infrastructure — benchmarks, evaluation protocols, educational resources, and community norms — that makes their work possible.

There is real urgency here. The bottlenecks to general-purpose robotics look breakable on a timeline of years, not decades, and the community that can make these systems safe must exist before they are widely deployed. The safety field has spent a decade building the tools to align systems that think. The next decade will deliver systems that think *and act*.

References

[1] Ilija Radosavovic, Sarthak Kamat, Trevor Darrell, and Jitendra Malik. Learning humanoid locomotion over challenging terrain, 2024. URL https://arxiv.org/abs/2410.03654.

[2] David Hoeller, Nikita Rudin, Dhionis Sako, and Marco Hutter. ANYmal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics*, 9(88), 2024. doi: 10.1126/scirobotics.adi7566. URL https://www.science.org/doi/abs/10.1126/scirobotics.adi7566.

[3] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS)*, 2024.

[4] Edgar Welte and Rania Rayyes. Interactive imitation learning for dexterous robotic manipulation: challenges and perspectives—a survey. *Frontiers in Robotics and AI*, 12, 2025.

[5] Kush Hari, Hansoul Kim, Will Panitch, Kishore Srinivas, Vincent Schorp, Karthik Dharmarajan, Shreya Ganti, Tara Sadjadpour, and Ken Goldberg. Stitch: Augmented dexterity for suture throws including thread coordination and handoffs. In *2024 International Symposium on Medical Robotics (ISMR)*. IEEE, 2024.

[6] Eric Jang, Dar Sleeper, and Brent Børnich. Neo home robot | Order today. https://www.1x.tech/discover/neo-home-robot, 2025. Accessed: 2026.

[7] Sunday Robotics. Beta program. https://www.sunday.ai/beta-program, 2026. Accessed: 2026.

[8] Dara Kerr. Google faces lawsuit after Gemini chatbot allegedly instructed man to kill himself. https://www.theguardian.com/technology/2026/mar/04/gemini-chatbot-google-jonathan-gavalas, 2026. Accessed: 2026.

[9] Coefficient Giving. Strategic cause selection. https://coefficientgiving.org/research/strategic-cause-selection/, 2025. Accessed: 2026.

[10] Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan

Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a VLA that learns from experience, 2025. URL https://arxiv.org/abs/2511.14759.

[11] Abhay Deshpande, Maya Guru, Rose Hendrix, Snehal Jauhri, Ainaz Eftekhar, Rohun Tripathi, Max Argus, Jordi Salvador, Haoquan Fang, Matthew Wallingford, Wilbert Pumacay, Yejin Kim, Quinn Pfeifer, Ying-Chun Lee, Piper Wolters, Omar Rayyan, Mingtong Zhang, Jiafei Duan, Karen Farley, Winson Han, Eli Vanderbilt, Dieter Fox, Ali Farhadi, Georgia Chalvatzaki, Dhruv Shah, and Ranjay Krishna. MolmoB0T: Large-scale simulation enables zero-shot manipulation, 2026. URL https://arxiv.org/abs/2603.16861.

[12] NVIDIA: Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

[13] Cynthia Breazeal. *Designing sociable robots*. MIT press, 2004.

[14] Linda B. Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2):13–29, 2005.

[15] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, 2026.

[16] Jared Perlo, Alexander Robey, Fazl Barez, Luciano Floridi, and Jakob Mökander. Embodied AI: Emerging risks and opportunities for policy action, 2025. URL https://arxiv.org/abs/2509.00117.

[17] 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 *International Conference on Learning Representations*, volume 2024, pages 30988–31043, 2024.

[18] Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study, 2025. URL https://arxiv.org/abs/2503.17473.

[19] Brian D. Earp, Sebastian Porsdam Mann, Mateo Aboy, Edmond Awad, Monika Betzler, Marietjie Botes, Rachel Calcott, Mina Caraccio, Nick Chater, Mark Coeckelbergh, Mihaela Constantinescu, Hossein Dabbagh, Kate Devlin, Xiaojun Ding, Vilius Dranseika, Jim A. C. Everett, Ruiping Fan, Faisal Feroz, Kathryn B. Francis, Cindy Friedman, Orsolya Friedrich, Iason Gabriel, Ivar Hannikainen, Julie Hellmann, Arasj Khodadade Jahrome, Niranjan S. Janardhanan, Paul Jurcys, Andreas Kappes, Maryam Ali Khan, Gordon Kraft-Todd, Maximilian Kroner Dale, Simon M. Laham, Benjamin Lange, Muriel Leuenberger, Jonathan Lewis, Peng Liu, David M. Lyreskog, Matthijs Maas, John McMillan, Emilian Mihailov, Timo Minssen, Joshua Teperowski Monrad, Kathryn Muyskens, Simon Myers, Sven Nyholm, Alexa M. Owen, Anna Puzio, Christopher Register, Madeline G. Reinecke, Adam Safron, Henry Shevlin, Hayate Shimizu, Peter V. Treit, Cristina Voinea, Karen Yan, Anda Zahiu, Renwen Zhang, Hazem Zohny, Walter Sinnott-Armstrong, Ilina Singh, Julian Savulescu, and Margaret S. Clark. Relational norms for human-AI cooperation, 2025. URL https://arxiv.org/abs/2502.12102.

[20] Anton Korinek and Joseph E Stiglitz. Artificial intelligence and its implications for income distribution and unemployment. In *The economics of artificial intelligence: An agenda*, pages 349–390. University of Chicago Press, 2018.

[21] Andrew Berg, Edward F Buffie, and Luis-Felipe Zanna. Should we fear the robot revolution? (the correct answer is yes). *Journal of Monetary Economics*, 97:117–148, 2018.

[22] John Danaher and Sven Nyholm. Automation, work and the achievement gap. *AI and Ethics*, 1(3):227–237, 2021.

[23] Anca Gheaus and Lisa Herzog. The goods of work (other than money!). *Journal of Social Philosophy*, 47(1), 2016.

[24] Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental AI development, 2025. URL https://arxiv.org/abs/2501.16946.

[25] David Gray Widder, Meredith Whittaker, and Sarah Myers West. Why 'open' AI systems are actually closed, and why this matters. *Nature*, 635(8040):827–833, 2024.

[26] Callum Sharrock, Lukas Petersson, Hanna Petersson, Axel Backlund, Axel Wennström, Kristoffer Nordström, and Elias Aronsson. Butter-bench: Evaluating LLM controlled robots for practical intelligence, 2025. URL https://arxiv.org/abs/2510.21860.

[27] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025.

[29] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models, 2025. URL https://arxiv.org/abs/2501.09747.

[30] Bear Häon, Kaylene Caswell Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models. In *Conference on Robot Learning*, pages 2743–2762. PMLR, 2025.

[31] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

[32] TRI LBM Team, Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, Naveen Kuppuswamy, Kuan-Hui Lee, Katherine Liu, Dale McConachie, Ian McMahon, Haruki Nishimura, Calder Phillips-Grafflin, Charles Richter, Paarth Shah, Krishnan Srinivasan, Blake Wulfe, Chen Xu, Mengchao Zhang, Alex Alspach, Maya Angeles, Kushal Arora, Vitor Campagnolo Guizilini, Alejandro Castro, Dian Chen, Ting-Sheng Chu, Sam Creasey, Sean Curtis, Richard Denitto, Emma Dixon, Eric Dusel, Matthew Ferreira, Aimee Goncalves, Grant Gould, Damrong Guoy, Swati Gupta, Xuchen Han, Kyle Hatch, Brendan Hathaway, Allison Henry, Hillel Hochsztein, Phoebe Horgan, Shun Iwase, Donovon Jackson, Siddharth Karamcheti, Sedrick Keh, Joseph Masterjohn, Jean Mercat, Patrick Miller, Paul Mitiguy, Tony Nguyen, Jeremy Nimmer, Yuki Noguchi, Reko Ong, Aykut Onol, Owen Pfannenstiehl, Richard Poyner, Leticia Priebe Mendes Rocha, Gordon Richardson, Christopher Rodriguez, Derick Seale, Michael Sherman, Mariah Smith-Jones, David Tago, Pavel Tokmakov, Matthew Tran, Basile Van Hoorick, Igor Vasiljevic, Sergey Zakharov, Mark Zolotas, Rares Ambrus, Kerri Fetzer-Borelli, Benjamin Burchfiel, Hadas Kress-Gazit, Siyuan Feng, Stacie Ford, and Russ Tedrake. A careful examination of large behavior models for multitask dexterous manipulation, 2025. URL https://arxiv.org/abs/2507.05331.

[33] Hugo Buurmeijer, Carmen Amo Alonso, Aiden Swann, and Marco Pavone. Observing and controlling features in vision-language-action models, 2026. URL https://arxiv.org/abs/2603.05487.

[34] Aiden Swann, Lachlain McGranahan, Hugo Buurmeijer, Monroe Kennedy III, and Mac Schwager. Sparse autoencoders reveal interpretable and steerable features in VLA models, 2026. URL https://arxiv.org/abs/2603.19183.

[35] Brett Adcock. Roadmap to a positive future powered by AI. https://www.figure.ai/master-plan, 2022. Accessed: 2026.

[36] Bureau of Labor Statistics. The employment situation - March 2026. https://www.bls.gov/news.release/pdf/empsit.pdf, 2026. Accessed: 2026.

[37] Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Huazuo Gao, Jiashi Li, Liyue Zhang, Panpan Huang, Shangyan Zhou, Shirong Ma, Wenfeng Liang, Ying He, Yuqing Wang, Yuxuan Liu, and Y. X. Wei. Insights into DeepSeek-V3: Scaling challenges and

reflections on hardware for AI architectures, 2025. URL https://arxiv.org/abs/2505.09343.

[38] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J Pappas. Jailbreaking LLM-controlled robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11948–11956. IEEE, 2025.

[39] Kensuke Nakamura, Lasse Peters, and Andrea Bajcsy. Generalizing safety beyond collision avoidance via latent-space reachability analysis. In *Robotics Science Systems*, 2025.

[40] Ravi Pandya, Madison Bland, Duy P Nguyen, Changliu Liu, Jaime Fernández Fisac, and Andrea Bajcsy. From refusal to recovery: A control-theoretic approach to generative AI guardrails. In *Proceedings of IASEAI Conference*, volume 2, pages 513–526, 2026.

[41] Rhitu Chatterjee. Their teenage sons died by suicide. now, they are sounding an alarm about AI chatbots. https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide, 2025. Accessed: 2026.

[42] Troy Griggs and Daisuke Wakabayashi. How a self-driving Uber killed a pedestrian in Arizona. https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html, 2018. Accessed: 2026.

[43] Talia Soglin and Alice Yin. Food delivery robots shatter two Chicago bus shelters. "Two in seven days is not great," alderman says. https://www.chicagotribune.com/2026/03/26/food-delivery-robots-bus-shelters-chicago/, 2026. Accessed: 2026.

[44] Pierre Sermanet, Anirudha Majumdar, Alex Irpan, Dmitry Kalashnikov, and Vikas Sindhwani. Generating robot constitutions benchmarks for semantic safety, 2025. URL https://arxiv.org/abs/2503.08663.

[45] MATS. MATS: ML alignment theory scholars program. https://www.matsprogram.org, 2026. Accessed: 2026.

[46] FAR.AI. FAR.AI events. https://www.far.ai/events, 2026. Accessed: 2026.

[47] Center for AI Safety. Introduction to ML safety. https://course.mlsafety.org/about, 2023. Accessed: 2026.