

Simulation Operators: The Next Level of the Annotation Business

Ardy Haroen
University of Southern California

haroen@usc.edu

We bet on agentic AI being integrated into other domains within the next few years: healthcare, manufacturing, automotive, etc., and the way it would be integrated is into cyber-physical systems, which are systems that integrate the computer brain into a physical receptor/actuator (e.g. robots). As the demand for cyber-physical agents increases, so would be **the need to train and align them**.

We also bet on the scenario where frontier AI and robotics labs would not be able to handle all of the demands for training and aligning those agents, especially in specific domains, therefore leaving opportunities for other players to fulfill the requirements for training those agents: **providing a dataset of scenarios to fine-tune the agents and providing the people to give feedback to the model for alignment**.

Furthermore, we also bet that human intervention would still be required to supervise deployed agents, as demanded by various regulations. Therefore leaving opportunities to **develop supervision platforms** which might highly differ between different industries.

Our writeup below is an attempt to analyze what's required to close those gaps.

1. Problem overview

Below we elaborate the arguments for our bets:

Bet #1: Increase of AI Integration in the Form of Cyber-Physical Systems

- Indicated by the [AIAT Report](#), various industries demand AI. And there are many industries that cannot benefit from AI without connecting them to physical systems, and these systems would be safety-critical. E.g.: Self-driving cars, AI-assisted personalized medicine, AI which patrols a factory to ensure safety.
- Frontier AIs that we have now are still digital entities that cannot take action on their own.

Bet #2: Frontier AI/Robotic Labs Would Not be Able to Handle All Training and Alignment Demands

- Clients from different industries might need highly differing requirements, and even in one industry there might be highly different use cases. E.g: Self-driving cars are a highly different domain with automotive manufacturing. An AI agent for automatically handling a vehicle and an AI agent for assembling parts of a car are trained differently and evaluated

differently. Even if we have AGI, it needs to recognize different concepts and comply with different sets of regulations for each task.

- One concrete example of this is Boston Dynamics. Even as a “frontier robotics lab”, it still [forms partnerships](#) with other companies to fulfill its clients’ needs.
- We still need humans to provide feedback when training the algorithm, and domain specific examples might need domain-experts to give feedback. Frontier AI Labs might not have the capacity to do the hiring and manpower management. Besides that, human oversight is also mandated by regulations such as the [EU AI Act](#). By this argument, even if we are progressing towards scalable oversight, I argue that business owners would not trust such algorithms easily, and therefore we should not build startups assuming that AI oversight is easily accessible.

Bet #3: Increase in Demand for Supervision Platforms

- Regulations mandate human supervision for high-risk AI systems ([GDPR](#), [EU AI Act](#))
- Same as above, different use cases might require highly different supervision platforms, as each platform might need its own human interface

2. Your solution

Platform Features

- **Annotation and Simulation:** Promptable simulator with editing features.
 - As an annotator, user could define the requirements of the environment of the business (e.g. the specifics of the factory)
 - As an annotator, user could utilize GenAI to simulate scenarios for which the agent could act in the form of video, so that the scenarios could be used to train the AI.
 - The scenarios should include various adversarial scenarios so that the user could gain a better understanding of how the agent could go wrong, and so that the models could be more robust to adversarial conditions (e.g. unexpected weather, unexpected temperature conditions, attempts on jailbreaking, etc.
 - As an annotator, user can edit the scenarios that the GenAI produced and annotate different types of information to the scenarios produced
- **Annotator/Evaluator Hiring Platform**
 - As the business owner, user could recruit people with expertise on the domain in which the agent would operate
 - As a domain expert, user could apply to various annotation and evaluation jobs in the platform
- **Supervision Platform**
 - As an evaluator/supervisor, user could determine what actions the agent should take when it arrives at a critical point in which human intervention/human judgement is needed

- As an evaluator/supervisor, user could put a stop to the agent should the user determine the agent has malfunctioned in any way (misaligned, stuck in an erratic behavior, etc.)

Would This Product be In Demand?

In the appendix below (**Indications of Demand**), we listed indicators that this product would have a product-market-fit.

Is This Technically Feasible? – Current State of Research

- **RLHF:** This would be one of the baseline techniques we would implement. The problem with RLHF is how we can transform the input and output compared to the original paper, as the original RLHF paper was intended for textual input and output. While industrial use cases might require different types of input (e.g. Pressure sensor readings, temperature) different output (e.g. Sequence of action to fix a pipe)
- **UniSim:** Google has developed an PoC for a video-generation model which outputs sequences of action and uses those generated videos to train their robotic arm to perform general tasks. Along with other video-generation AI systems (e.g. Sora) and displays in robotics (e.g. Figure) I believe we could generate more varying samples and apply them to the varying industries

3. Pilot experiment or demo

Omitted – we focused more on problem validation

4. Process

Timeframe	What will you do?
Next 3 months	A simple MVP would be to build the gig economy network: ● Work with benchmark and eval companies e.g. \vectorview, METR ● Produce new industry benchmarks (e.g. for QnA RAG, for healthcare assistance, etc.) – with jailbreaking attempts included in the benchmarks, as an attempt to improve the current GenAI adaptation landscape in the industry. ● Test the water – talk with many people from different companies to see which ones would really need our product
2025	● Launch the universal simulation platform and its corresponding data annotation platform – conduct extensive research on generating simulations ● Secured partnership with at least 3 companies – Started working on their use case ● Identified which key industries would be on high

	priority to be secured
2026	• Develop the human-in-the-loop supervision platform for deployed AI agents
2027	• Launch the supervision platform

5. Impact on AI safety & key risks

Positive Impact on AI Safety

- **Accelerating Domain-Specific Alignment:** Alignment in many domains cannot wait for a safe AGI, we would need to start small and iterate on alignment within various domains
- **Creation of Jobs:** One of the fears of many regulating bodies regarding AI progress is the loss of jobs. This product promotes a job market which would hopefully lighten the transition from human to AI operators

Risks on AI Safety

- **Capabilities Externalities:** This will certainly accelerate the technical progress and agency that some might fear. E.g. to accelerate the adaptation of AI agents to industrial domains
 - But we will need to focus on the human supervision part of this product which would enable alignment in those adaptations

6. Appendix

Introduction

Hi I'm Ardy. I was born and raised in Indonesia. I am currently pursuing my Masters in Computer Science @University of Southern California. Before my masters I worked as an AI Engineer for 3 years in one of the earliest e-commerce companies in Indonesia. I was fortunate enough to be one of the early people to engage with AI safety in Indonesia. Back in 2022, I participated in the Machine Learning Safety Scholars program as the only person from my country, and after that I became one of the first 10 members of EA Indonesia and built an AI safety reading group there. I'm currently planning to upskill in my masters while looking for the best way to make a positive impact (Currently I think it's by building my own company). I'm very much open for feedback and collaboration requests! I highly desire to give more opportunities to the people of my country to make a positive impact (Primarily in AI safety as this is my background). I believe there are many smart people there who just haven't been aware of such opportunities yet.

Indications of Demand

- **Job listings.** One of our vision is for our product to support the individuals whose work is listed as below:
 - Figure (robotics)

- <https://job-boards.greenhouse.io/figureai/jobs/4031371006>
 - <https://job-boards.greenhouse.io/figureai/jobs/4345940006>
- Mercor
 - https://docs.google.com/document/d/1ci0b012u45gL_JaNyckmqQO_KQ6qoZelFrbKpFjypzA/edit#heading=h.1w1s8su34zi5
- Boston Dynamics
 - https://bostondynamics.wd1.myworkdayjobs.com/Boston_Dynamics/job/Waltham-Office-POST/Simulation-Technical-Manager_R1454
 - https://bostondynamics.wd1.myworkdayjobs.com/Boston_Dynamics/job/Waltham-Office-POST/Simulation-Platform-Lead--Architect_R1422
- Neuralink
 - https://neuralink.com/careers/apply/?gh_jid=6098815003&gh_src=c356a2533us
- NVIDIA
 - DriveSim: Vehicle driving simulator
https://nvidia.wd5.myworkdayjobs.com/en-US/NVIDIAExternalCareerSite/job/China-Shanghai/Senior-AI-Engineer--DriveSim-Synthetic-Scenario-Generation_JR1980376?q=simulation
- **Boston Dynamics Use Cases.** These are sample :
 - Pharmacy Manufacturing:
<https://bostondynamics.com/case-studies/novo-nordisk-explores-spots-role-in-pharmaceutical-production/>
 - Automotive Manufacturing:
 - Whitepaper
<https://resources.bostondynamics.com/hubfs/Content%20Files/boston-dynamics-robots-in-automotive-manufacturing-whitepaper.pdf>
 - BMW
<https://bostondynamics.com/case-studies/spot-inspires-innovation-at-bmw-group-plant/>
 - Food Industry
 - Whitepaper
<https://resources.bostondynamics.com/hubfs/Content%20Files/boston-dynamics-food-beverage-whitepaper.pdf>

Key Doubts

- Would people want it?
 - Would the general people want it → yes they want jobs
 - Would companies want it?

- Probably. If there are cheaper options then they'd probably pick that
 - Probably. If they can build it themselves they'd probably do that
 - Probably not if they don't have that problem → need to really validate that they have that problem
- How far until this is economically valuable?
 - Now even research is still far from implementing agentic AI (e.g. Figure –the robotic company still puts 2yr as their expectation on going to market).
 - What I've seen from robotics companies' use cases, they're still developing one robot while trying to prove its usefulness to other industries, there's no one specific industry use case where it really clicked and innovation and funding is greatly flowing there.
- How will this improve capabilities along with safety?
 - This will certainly accelerate the technical progress that some might fear. E.g. to accelerate the adaptation of AI agents to industrial domains
 - But we will need to focus on the human supervision part of this product to make it not so safetywashing