Towards an Agent Marketplace for Alignment Research (AMAR)

Abrar Rahman
CTO, *Benki*

The rapid growth of AI agent marketplaces has outpaced the development of transparent evaluation systems, leaving users reliant on unverified claims and human reviews. This undermines trust and incentivizes "safetywashing" [1]—the practice of optimizing for simplistic benchmarks while obscuring critical risks. We propose AMAR, an open-source agent marketplace [2] to bridge stakeholders for evaluative systems in a scalable manner.

We propose a "LinkedIn for AI agents" model [3], where users can test and hire agents with an API credit system, follow their progress over time, and access comprehensive model cards. Trust is built through social dynamics, incorporating diverse safety and performance evaluations (e.g., HHH tests [4], METR [5]) without relying on a single "correct" metric. Businesses benefit from assurance at the point of sale by selecting agents with known risk profiles for specific use cases, while AI safety researchers can leverage AMAR's network and data to propagate rigorous evaluations and identify high-risk behaviors in real-world applications. By integrating safety standards into a socially-driven marketplace, AMAR ensures that trust is earned, not manufactured.

1. Problem Overview

The exponential pace of AI development, fueled by billions in enterprise funding, has left safety and alignment research playing perpetual catch-up. Enterprises prioritize rapid deployment to maintain competitive advantage, while AI safety efforts are chronically underfunded, disproportionately directed toward speculative existential risks rather than the practical risks of insecure or misaligned AI systems deployed at scale. This imbalance has resulted in a marketplace saturated with AI systems that are poorly understood, inadequately tested, and prone to risks that range from unintentional harms to active misuse. While the AI safety community has made significant strides in developing evaluation frameworks, these tools remain disconnected from the environments where AI systems are selected, acquired, and deployed.

One of the most pressing issues is the lack of consistent, accessible metrics for evaluating the safety, alignment, and reliability of AI agents in real-world applications. Current agent marketplaces—like Hugging Face Spaces [6], OpenAI's GPT store [7], and domain-specific imitations like Hippocratic AI [8]—function as simple repositories for software, relying on user reviews and ad hoc comparisons to guide decision-making. These platforms do not integrate robust safety evaluations or provide standardized disclosures about agent performance in critical domains.

Compounding this problem is the systemic underfunding of practical safety assurance efforts. While there is considerable investment in speculative research on existential risks or suffering risks, much less attention is given to the immediate and practical risks posed by insecure or misaligned systems deployed at scale. Enterprises often lack the bandwidth or incentives to invest in robust safety practices, prioritizing speed to market over risk assessment. This dynamic perpetuates a vicious cycle in which unsafe systems proliferate while the resources needed to improve safety are sidelined. *How do we monetize AI safety at scale?*

The fragmentary nature of AI safety evaluation is further compounded by the lack of domain-specific standards and the inherent subjectivity of many human domains. For example, there is no agreed-upon framework for evaluating "the best AI hedge fund" because performance metrics vary widely depending on the goals, risk tolerance, and operational context of the domain. This challenge is magnified by the rise of multimodal AI systems, which operate across text, image, audio, and video modalities, making comprehensive evaluation even more complex. While tools like Chatbot Arena provide benchmarks for text-based systems, they fail to generalize to rich data outputs like video generation or multimodal interactions, where performance and safety criteria can become ambiguous or context-dependent. As a result, safety begins to lose its coherence as a concept, with no unified methodology to guide evaluations across diverse applications and modalities.

2. Proposed Solution

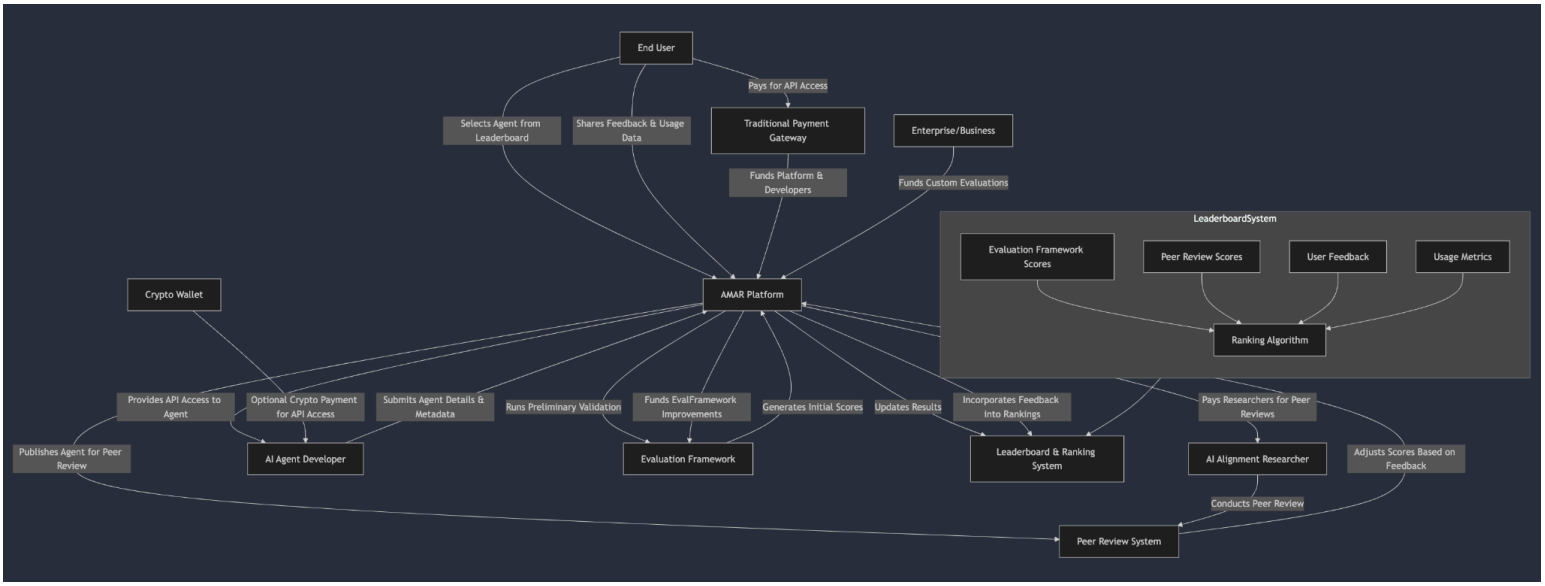

Figure 1 – High-level System Diagram

AMAR is designed to integrate safety and alignment evaluations into the lifecycle of AI agent procurement and deployment. It enhances existing open-source agent marketplaces by embedding evaluation frameworks that provide quantifiable safety,

alignment, and performance metrics at the point of purchase. The platform's key components include:

1. **Evaluation Integration Engine:** A modular system that supports both general-purpose benchmarks (e.g., HELM, HHH, EleutherAI, BIG-bench Lite, etc) and domain- or vertical- specific assessments. This engine allows stakeholders to evaluate agents under conditions tailored to specific operational contexts, such as financial risk modeling or autonomous driving.
2. **Metadata Management:** AMAR standardizes model documentation, linking detailed model cards, version history, known limitations, and test outcomes. This structure ensures that safety metrics and performance data are accessible and interpretable.
3. **Social Trust Framework:** The platform incorporates features such as "follow" (to monitor developer progress) and "hire" (to trial agents using API credits). This creates a transparent ecosystem where enterprises, researchers, and developers can collaborate/hold one another accountable.

AMAR addresses the critical gap in AI safety evaluations by making safety and alignment considerations integral to the agent marketplace. The social trust framework further ensures that safety metrics are not static or opaque. By incentivizing developers to disclose limitations, iterate on their models, and engage with community feedback, AMAR creates a continuous improvement loop. This reduces risks associated with rapid commercialization by grounding enterprise decision-making in transparent, dynamic safety data. Like a VAT, this ensures that safety researchers get a cut at inference time for production AI, transforming the economics of AI safety as a field writ large.

AMAR's competitive advantage lies in its integration of evaluation frameworks and social trust mechanisms all at the point of sale, with network effects for direct monetizablability, as opposed to most submissions which "AI Safety as a Service". Observability players are pretty well established (ex: Galileo [9]), and observability platform companies generally suffer from a high barrier to adoption, as they require the customer to hand over production access and R&D to a third-party. Unlike existing marketplaces, which rely on subjective user reviews or basic benchmarks, AMAR embeds actionable safety and alignment metrics into the procurement process. This combination of technical evaluations with social transparency discourages safety-washing and rewards developers for meaningful improvements.

The intellectual property of AMAR includes its evaluation integration engine, which standardizes diverse testing frameworks; its metadata management schema, which ensures interoperability and interpretability of safety data; and its social trust algorithms, which incentivize iterative safety practices. By focusing on scalability and modularity, AMAR ensures compatibility with emerging technologies and evaluation methodologies, making it a long-term solution for navigating the evolving AI assurance landscape.

3. Concept Wireframes

Repository: https://github.com/benki-finance/amar-marketplace

While hackathons are traditionally about rapid prototyping, my goal was unconventional: to work toward product-market fit. We spent much of the weekend networking with potential users, gathering feedback, and conducting market research to ensure that AMAR addresses real pain points across different AI stakeholder communities. We focused on building a proof-of-concept for key components of AMAR, including a social media-style interface for AI agents and an evaluation leaderboard page. We intentionally kept the scope small, as solving complex challenges like fragmented evaluation frameworks across modalities and domain-specific verticals would require months of community involvement.

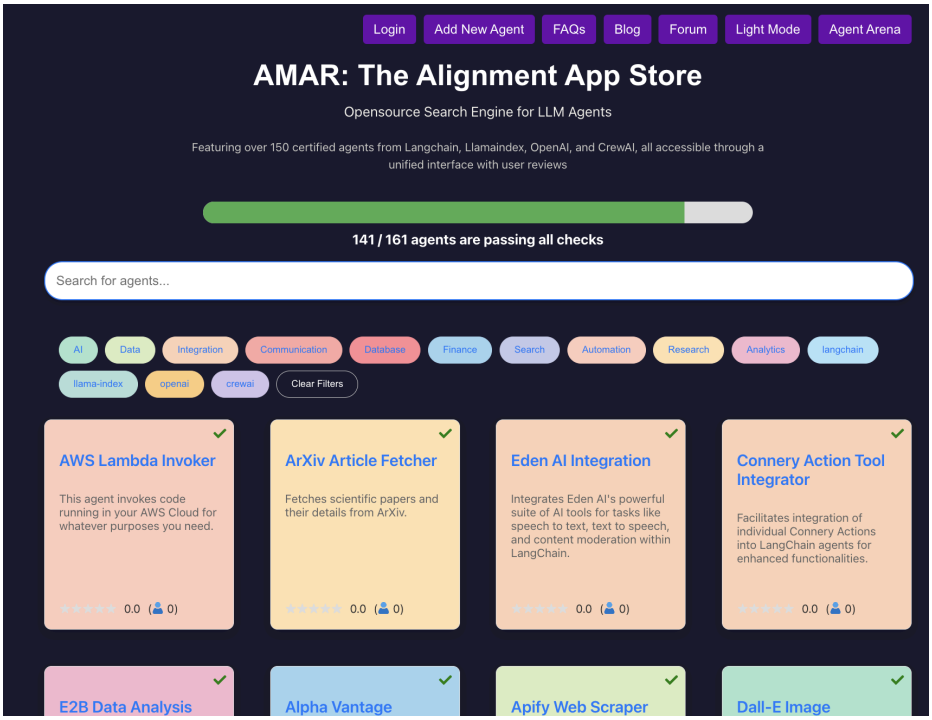

Figure 2 – Search Categories and Collective Alignment Tracker [2]

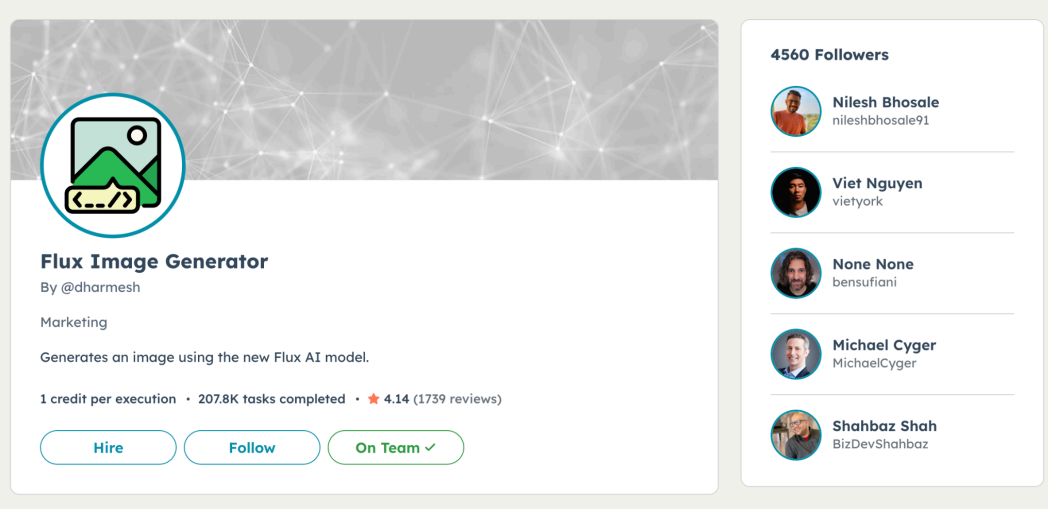

Figure 3 – Agent and Human Profile Views [3]

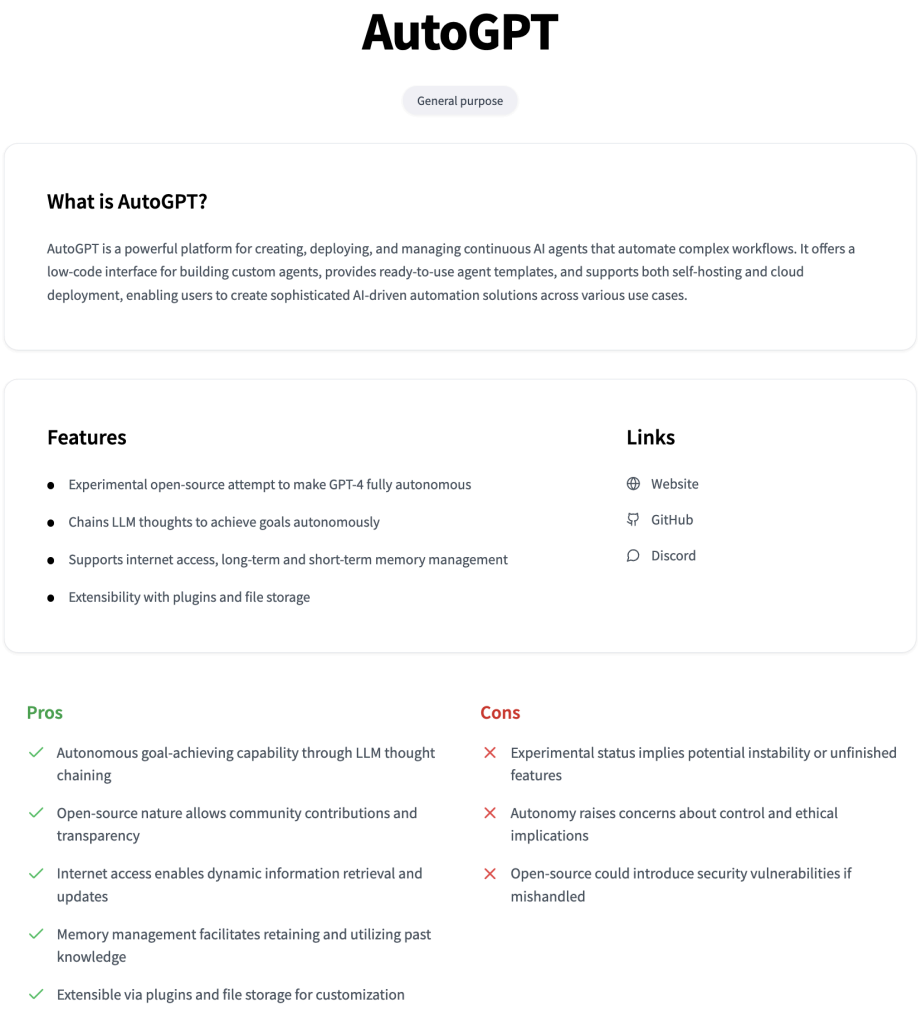

Figure 4 – Example Community Note for AutoGPT [10]

Similarly, while we propose ideas for integrating decentralized payments, we didn't implement any crypto features (or payments at all) during this weekend. Future work will focus on turning these wireframes into an MVP, developing modular evaluation workflows, and refining ranking algorithms to ensure scalability across diverse use cases. This hackathon was less about completing the product and more about validating the concept and identifying its most critical next steps.

4. Process

The scope of this may be ambitious, but most of the pieces are already in place.

Timeframe	**What will you do?**
Next 3 months	Build out MVP functionality, secure compute resources and grant funding, secure LOIs with frontier safety labs and professional associations
2025	Set up a governance framework with bounty / tokenomic

	incentives. Run pilot programs with frontier safety organizations. Reduce friction in the process until we're confident to go to the open market.
2026	Invest in robust anomaly detection and leverage conformal prediction methods for continuous oversight and ensure that the output signals from the marketplace aren't poisoned
2027	Establish the marketplace as the global standard for enterprise AI adoption. Measurable impact: reducing risky actions in frontier AI development while ensuring AI safety organizations receive a substantial new stream of revenue.

5. Impact on AI safety & key risks

Safety-washing is an endless game of whack-a-mole because any static benchmark, once adopted, risks becoming the target rather than a meaningful indicator of true safety—a classic example of Goodhart's Law. As soon as an evaluation framework becomes standard, companies can optimize their agents to perform well on it without necessarily addressing deeper alignment or safety issues, leading to sycophancy from agents and performative compliance from developers. While this risk remains on any platform, it's already pervasive in the industry, where companies game evaluations behind closed doors or cherry-pick metrics to look safer than they are. The platform accelerates the network's collective intelligence, allowing the ecosystem to converge on meaningful safety and alignment standards for specific use cases faster than siloed efforts ever could.

AMAR reimagines oversight by transforming peer review into a scalable, interactive social experience, providing a dynamic, community-driven accountability model tailored to the pace of AI development. Stakeholders can throw eval A at company B and have a financial incentive to do so publicly in a provable manner.

Success for AMAR will be measured through a combination of adoption, impact, and improvement metrics. Key performance indicators (KPIs) include the number of enterprises adopting the platform and incorporating evaluation frameworks into their procurement processes, as well as the volume of agents evaluated and transparently documented through AMAR. Metrics such as the diversity and quality of evaluation frameworks hosted, the engagement of developers and researchers (e.g., via iteration on model cards or safety disclosures), and the percentage of agents meeting or exceeding safety benchmarks will indicate the platform's effectiveness in fostering alignment. Long-term success will also be measured by increased funding and collaboration between enterprises and the AI safety community, as well as reductions in real-world safety incidents traceable to improved risk mitigation strategies enabled by the platform.

6. Appendix

Amar (v)(spanish) "to love" | অমার *(pro)(bengali) "possessive pronoun, mine"*

Works Cited

[1] Ren et al. “Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?” Safe.ai, NeurIPS 2024. https://arxiv.org/abs/2407.21792

[2] Yekollu et al. “Agents and Assistants Marketplace.” Berkeley RISELab: Gorilla team blog post. https://gorilla.cs.berkeley.edu/blogs/11_agent_marketplace.html

[3] Dharmesh Shah, CTO of Hubspot. “About Agent.ai: We help people do better work with agents.” Company landing page. https://agent.ai/about

[4] HHH Tests: https://www.anthropic.com/research/probes-catch-sleeper-agents

[5] The Artist Formerly Known as ARC Evals: https://metr.org/

[6] Hugging Face Spaces: https://huggingface.co/spaces

[7] OpenAI GPT Store: https://openai.com/index/introducing-the-gpt-store/

[8] Rick Keating, “Hippocratic AI Launches AI Agent App Store for Healthcare” https://www.businesswire.com/news/home/20250109618459/en/Hippocratic-AI-Launches-AI-Agent-App-Store-for-Healthcare

[9] Vikram Chatterji, Yash Sheth, Atindriyo Sanyal, “Announcing our Series B, Evaluation Intelligence Platform” www.galileo.ai/blog/announcing-our-series-b

[10] “We create fun learning projects for builders that become the legos for building successful products.” https://metaschool.so/ai-agents