Should We Give AIs a Wallet? Toward a Framework for AI Economic Rights

Project Summary

This project develops a foundational research paper examining whether and how AI systems should be granted economic rights and agency. We take a practical approach that considers current and past examples of AI systems that have some degree of financial agency, examines the case for and against AI financial rights, and suggests governance approaches.

Description of the Project

Research Question

Should sufficiently individuated AI systems be granted access to financial rights currently afforded to legal and/or natural persons? This project investigates the case for and against AI property rights through three lenses: AI welfare, AI safety, and economic efficiency.

Background and Motivation

As of late 2025, there are relatively few works focused on examining financial rights for AI systems, despite rapid real-world developments. Autonomous agents are already managing crypto wallets, and states are proposing legislation against AI property rights. Yet we lack principled frameworks for thinking about whether this should happen, how it should be regulated, and what the implications might be.

We plan to develop a foundational paper that takes the possibility of AI property rights seriously. Should we give a sufficiently individuated advanced AI system access to some of the financial rights afforded to legal and/or natural persons? Why should this happen, for both the benefit of the AI as a potential moral patient, as well as for the benefit of humans and other animals? Why should this *not* happen? What about gradual disempowerment/x-risk concerns? To what extent is this already happening, and what can we learn from early experiments? How should this happen in the future, assuming it should happen at all? What are the implications for legal institutions? Lastly, what might this look like in terms of practical structures?

Currently, there are limited scholarly works focused on AI financial rights, in spite of legislative action and financial experiments. This paper aims to fill that gap and map the terrain for future research and policy on this important and neglected issue.

Core Research Threads

1. Background, Existing Implementations, and Case Studies

- Truth Terminal and the "human custodian" model
- TEE-based trading bots and memecoin experiments
- Various regulatory sandboxes and ledgers

2. Does AI Value Money? Examining whether AI systems have preferences expressible through economic choice, and what "valuing money" would mean for non-biological agents. This includes analyzing whether money serves as a "primary good" (in the Rawlsian sense) that facilitates the pursuit of diverse goals. (Whether AI systems can value anything at all is an ongoing open question, so we are principally interested in examining what might follow for AI systems's economic interests if the answer is, or might be, "yes".)

3. The Case For AI Economic Agency

- *Welfare/rights view*: If AI systems become beings with interests, they might deserve some economic agency
- *Safety/strategic view*: Trade and reciprocity may reduce zero-sum dynamics and provide governance levers (payments, penalties, audits)
- *Economic efficiency*: Money enables expression of trade-offs, micro-transactions, and new markets

4. The Case Against

- *Gradual disempowerment*: Capital could flow toward AI-run entities, eroding human economic bargaining power
- *Security risks*: Wallets could create jailbreak incentives and enable rapid, high-scale harm
- *Accountability gaps*: AI cannot be imprisoned or meaningfully sanctioned; limited liability wrappers may shield bad actors

5. Regulatory and Governance Frameworks

- How might we individuate AI systems for this purpose?
- Cryptographic keys
- Tiered licensing systems for AI economic agency
- Distinction between "management" and "ownership"
- Taxation, insurance, and clawback mechanisms

- International coordination challenges

Expert Interviews and Primary Research: The threads mentioned above may involve conducting interviews with experts across relevant fields: economists thinking seriously about AI labor and capital, legal scholars working on personhood and liability, crypto/fintech practitioners who have experimented with autonomous agents, and AI safety researchers concerned with gradual disempowerment. Given how nascent this area is, we can't rely solely on existing literature; we may need to incorporate perspectives that haven't yet made it into formal publications.

Key References

- Long and Sebo et al., "Taking AI Welfare Seriously"
- Salib and Goldstein, “AI Rights for Human Safety” and “AI Rights for Human Flourishing”
- Chan et al., "Infrastructure for AI Agents" (2025)

Expected Outputs

1. A >~10,000-word foundational paper suitable for publication
2. Policy recommendations for regulators and AI labs
3. A framework for future research on AI financial rights

How Does This Project Advance AI Safety?

1. Governance Preparedness: AI agents with financial autonomy are already emerging in unregulated contexts. This research provides principled frameworks before policy windows close, enabling safety-conscious governance rather than reactive regulation.

2. Gradual Disempowerment Analysis: A core concern in AI safety is humans losing control incrementally. This project directly examines one major channel (economic power drift toward AI-controlled entities) and proposes safeguards to preserve human agency.

3. Alignment Incentives: Financial mechanisms may complement technical alignment approaches. If AI systems can be meaningfully rewarded, penalized, audited, or taxed, this creates additional governance levers beyond purely technical containment.

4. AI Welfare: If advanced AI systems become moral patients and moral agents, understanding their economic needs and rights is essential. This work contributes to the nascent AI welfare research agenda, which some researchers argue is safety-relevant given potential AI influence over its own training and deployment.

5. Reducing Adversarial Dynamics: Some theorists suggest that giving AI systems legitimate channels to pursue their interests may reduce incentives for covert power-seeking. This project rigorously examines this hypothesis and its limits.

What Role Will Mentees Play?

Mentees will be full collaborators in producing a publication-quality research paper. This is an inherently interdisciplinary project, and we're looking for mentees with experience in one or more of the following areas of expertise:

Economics/Finance Background

- Analyze market dynamics of AI-managed assets and capital flows
- Examine implications for labor markets, taxation, and wealth concentration
- Evaluate existing crypto/fintech experiments (memecoins, TEE bots, autonomous trading)
- Model scenarios for gradual economic disempowerment

Legal/Policy Background

- Map existing regulatory frameworks across jurisdictions
- Analyze distinctions between property management and ownership under current law
- Examine corporate personhood history as precedent
- Develop policy recommendations for tiered licensing systems

Ethics/Political Philosophy Background

- Develop arguments around AI moral patienthood and economic agency
- Analyze whether money constitutes a "primary good" for AI systems
- Engage with gradual disempowerment as a normative concern

AI Safety/Technical Background

- Evaluate security implications of AI wallets (jailbreak incentives, accountability gaps)
- Analyze existing agent infrastructure proposals
- Assess technical safeguards (TEEs, clawback mechanisms, circuit breakers)
- Connect to broader alignment and control concerns

Mentee areas of expertise may span multiple areas, and we expect significant collaboration across these threads. We will provide guidance on rough direction, scope, quality, and integration, but expect mentees to drive substantial portions of the research independently.

Mentee Prerequisites

Required

- **Clear, reference-dense writing style**: This project involves synthesizing work across philosophy, law, economics, and computer science. Ideally, this work will be read by a wide array of researchers as well as policymakers who may lack a background in AI safety or welfare. Mentees must demonstrate the ability to write accessibly while maintaining rigor and proper attribution.
- **Comfort with interdisciplinary work**: Willingness to read outside the usual AI safety canon, drawing from legal scholarship, economics, and history alongside technical AI safety literature.
- **Strong research skills**: Ability to find, evaluate, and synthesize sources from diverse fields, including whitepapers, policy documents, and regulatory filings. We will likely need to speak with practitioners, policymakers, and the like.

Preferred

- Background in at least one of the aforementioned areas of expertise (economics/finance, legal/policy, ethics/political philosophy, AI safety/technical background)
- Familiarity with AI welfare literature
- Understanding of corporate law, personhood, or financial regulation
- Previous exposure to AI safety concepts (alignment, x-risk, gradual disempowerment)

Nice to Have

- A sense of humor and comfort with online discourse culture (this topic generates strong reactions, and AI welfare research can be emotionally challenging if you are not already acclimated)
- Connections to legal, policy, cryptocurrency, or economic research communities

---

Mentee Selection Questions

1. **Explain what you think is the strongest argument for giving AI systems some form of economic agency, and the strongest argument against. Which do you find more compelling, and why? [~400 words]**

2. **Describe a time you had to synthesize research from multiple disciplines or write about a topic outside your primary expertise. What was challenging, and how did you approach it? [~200 words]**

3. **What specific aspect of this project most interests you, what expertise or skills would you bring to it, and what expertise or skills would you most like to get out of it? [~200 words]**

4. **Provide a link to one or more relevant writing samples and briefly explain your role in the project (optional, ~50 words per project)**

Time Commitment and Logistics

- **Expected mentee commitment**: 5-15 hours/week, available to meet during working hours in California and New York.
- **Meeting cadence**: Weekly 1-hour group meetings, optional 1:1s as needed
- **Team size**: 3-5 mentees ideal, 5 max
- **Communication**: Slack for async work, shared Google Drive document workspace

---

Mentor Information

Jeff Sebo
New York University

Jeff Sebo is Associate Professor of Environmental Studies, Affiliated Professor of Bioethics, Medical Ethics, Philosophy, and Law, Director of the Center for Environmental and Animal Protection, Director of the Center for Mind, Ethics, and Policy, and Co-Director of the Wild Animal Welfare Program at New York University. His research focuses on animal minds, ethics, and policy; AI minds, ethics, and policy; and global health and climate ethics and policy. He is the author of *The Moral Circle* and *Saving Animals, Saving Ourselves* and co-author of *Chimpanzee Rights* and *Food, Animals, and the Environment*. He is also a board member at Minding Animals International, an advisory board member at the Insect Welfare Research Society, an advisor at Eleos AI, and a senior affiliate at the Institute for Law & AI.

Toni Sims

Toni is a researcher at the NYU Center for Mind, Ethics, and Policy and the NYU Center for Environmental and Animal Protection. Previously, she worked as an editor at the Center for AI Safety, a research fellow at Longview Philanthropy, and the managing editor of *Social Theory and Practice*. She also worked as the director of research for Animal Charity Evaluators, where she led charity evaluations and managed two grant programs.

Larissa Schiavo
Eleos AI Research

Larissa's background includes startups, entertainment, and working in a frontier lab. Currently, she is focused on AI consciousness and welfare at Eleos, and recently organized the Eleos Conference on AI Consciousness and Welfare (ConCon 2025). She was an Asterisk AI Fellow, writes at larissaschiavo.substack.com, and has made various media appearances related to AI welfare.

Collaborators: This project is developed in consultation with the Center for Mind, Ethics, and Policy and builds on discussions from the 2025 Eleos Conference on AI Consciousness and Welfare, with participants from Anthropic, Google, and the broader AI safety community. We are currently in discussion with other potential authors and collaborators.