From Asimov's Laws to Ethereum's Protocol: [Re]searching the intersection where crypto meets AI alignment.

[

1074×638 140 KB

](https://collective.flashbots.net/uploads/default/original/2X/1/1c9eb70663049a9fc1d9e2c864907d0fbe8a81d8.png)

An Unfinished Treasure Map

Both the cryptoeconomics research community and the AI safety / new cyber-governance / existential risk community are trying to tackle what is fundamentally the same problem: How can we regulate a very complex and very smart system with unpredictable emergent properties using a very simple and dumb system whose properties once created are inflexible.

- Vitalik Buterin, Why Cryptoeconomics and X-Risk Researchers Should Listen to Each Other More(2016)

Over the course of the past few years, there has been many explorations of combining crypto and AI in solving practical problems, e.g. MEV and commitment races, credible auctions via blockchains, conditional information disclosure and programmable privacy, federated learning, zkml, and identity. Recently, transactions on Ethereum are becoming more like agentic intents (e.g., account abstraction, smart transactions), and protocols like SUAVE arise to turn adversarial on-chain bot war into coordinated execution markets that satisfy human preferences.

Here in Zuzalu, we attempt to explore their intersection from first principles. During an evening whiteboarding session at a Pi-rate Ship

pop-up hackerhouse, we, a group of humans, started by brainstorming the core concepts that underpins the foundations of both fields. We arrived at a collective mindmap for Crypto X AI, taken inspiration from MEV mindmap: undirected traveling salesman. This exploration leads us to a continuing journey into a future where crypto mechanisms become increasingly conscious, and AI plays a transformative role in prediction and alignment.

[

2548×1492 443 KB

](https://collective.flashbots.net/uploads/default/original/2X/4/4d1733f948624c0d0e5cdae57a61bfc9937ce933.png)

This is just the beginning of our journey exploring the convergence of crypto and AI alignment research… If you would like to contribute to our collective mindmap, ping @sxysun1

on Twitter.

Agenda

Time:

11:00 - 19:30 (GMT+2) on Sunday, May 7, 2023

Location:

The Lighthouse, Zuzalu

Livestream:

zuzalu.streameth.org

Pre-game:

Wanna have your questions answered by the speakers? Want the event to focus on what you are interested in? [Make yourself heard

](https://www.notion.so/flashbots/CryptoXai-wtf-Zuzalu-Sunday-7th-9587c86ee7ed4ddeadbd862dc13a0c7f?pvs=4)!

Chapter I. Putting X-Risk in Perspectives

11:00-11:30 Daniel Kang: AI Capabilities and Our Greatest Fears: A Collective Timeline

(Slides) (Recording)

Abstract: & Background Reading

- Abstract:

Discussions of risks from AI are often emotionally charged, making it difficult to have a common understanding of what the concrete risks are and what capabilities will lead to these risks. In this talk, Daniel will provide an opinionated view on the history and future of these capabilities/risks. The goal will be to align the workshop attendees on concrete scenarios to discuss.

- Background Reading:
- A brief (opinionated) view on the history of capabilities
- An overall explainer of risks
- How the EU is thinking about risks
- A brief (opinionated) view on the history of capabilities
- An overall explainer of risks
- How the EU is thinking about risks

Add & vote on statements

11:30-12:00 Vitalik Buterin: AI Dystopias vs. AI Utopias

(Slides) (Recording)

Abstract

- Abstract:

Laying out hypothetical scenarios of dystopian and utopian outcomes of AI development on humans, and potential paths as to how to get there. * Dystopias: FOOM Doom, medium-speed descent into madness, locked in: totalitarianism, stagnation, civilizational decline, the human Moloch: what might cause the above

- Utopias: Fun theory, CEV and other theories of aggregating human preferences, “Minimal AI government” ideas, Human coordination: what might cause the above
- Dystopias: FOOM Doom, medium-speed descent into madness, locked in: totalitarianism, stagnation, civilizational decline, the human Moloch: what might cause the above
- Utopias: Fun theory, CEV and other theories of aggregating human preferences, “Minimal AI government” ideas, Human coordination: what might cause the above

Add & vote on questions!

12:00-12:30 Nikete: Eventually We All Die

(Slides) (Recording)

Abstract: & Background Reading

- Abstract:

Competing existential risks, falsifiability, near-misses, and sharing the planet with cognitively more advanced agents. With a side of LoRa, Markets and Predictions.

Understanding the relative sizes of x-risks is crucial to guiding policy. Increasing AI capabilities have two effects on this, increasing the probability that the AI destroys us at some point, while potentially also preventing things that destroy us. Credibly signalling who understands the relative size of these two counter-acting forces on human welfare is crucial for our future. This talk proposes an initial mechanism towards this in the form of fine-tunings of LLMs that reflect beliefs, and are judged on their ability to predict tommorows news given todays.

- Background readings:
- Decision Scoring Rules
- Decision Rules and Decision Markets
- The Singularity in Our Past Light-Cone
- Forecasting Future World Events with Neural Networks

- The A.I. Dilemma: Growth versus Existential Risk
- LoRA: Low-Rank Adaptation of Large Language Models
- Decision Scoring Rules
- Decision Rules and Decision Markets
- The Singularity in Our Past Light-Cone
- Forecasting Future World Events with Neural Networks
- The A.I. Dilemma: Growth versus Existential Risk
- LoRA: Low-Rank Adaptation of Large Language Models

Add & vote on questions!

12:30-13:00 Mike Johnson: The Limits of the Utility Function and the Structure of a New Science of Consciousness

(Slides) (Recording)

Abstract: & Background Reading

- Abstract:

Consciousness is a pre-scientific phenomenon similar to alchemy. What would it take to transition to a principled chemistry of phenomenology? This talk will survey “what kind of problem” consciousness is, what we might expect from a mature science of consciousness, and my solutions thus far. Implications for “what kind of thing humans are” and AI alignment will be discussed.

- Background readings:
- Symmetry of Valence (STV) Primer
- It From Bit
- Principia Qualia (condensed version)
- Principia Qualia (original version)
- Making and breaking symmetries in mind and life
- Symmetry of Valence (STV) Primer
- It From Bit
- Principia Qualia (condensed version)
- Principia Qualia (original version)
- Making and breaking symmetries in mind and life

Add & vote on questions!

13:00-14:00 Lunch Break(out) Session: Lightning Talks

5min lightning talks - topics collected from event attendees

- How to Overcome the Fear of AGI - Han
- Making a Positive singularity Incentive-compatible (Roko)
- Levels of Defence in AI Safety (A. Turchin)
- AI-Directed: Adding a New Cultural Type to D. Riesman’s the Lonely Crowd (R. Yager)

Chapter 2. A Rational Hitchhiker’s Guide to AI Alignment

This chapter aims to provide a survey of the various approaches to AI alignment.

14:00-14:30 Jessica Taylor: An Overview of AI Alignment Research, 2008-2023

(Slides) (Recording)

Abstract

Abstract:

What progress has the AI alignment field made over time? How have its problems been formulated, reframed, and solved over time? What are some of the fundamental obstacles? This talk presents an overview of the history of the field and current lines of inquiry.

Add & vote on questions!

14:30-15:00 Rob Knight, Nate Soares: Seaside Chat: Alignment as a Layer in the Stack

Abstract

Abstract:

Brief introduction to the idea of “perspective-taking”, trying to understand why someone has the opinion that they do

- Let’s try to separate out alignment from other concerns, e.g. ethics, by analogy with other layer/stack systems in computing:
- TCP/IP 5-layer model (stacked protocols)
- File system stack
- TCP/IP 5-layer model (stacked protocols)
- File system stack
- How does an approach like reinforcement learning fail to ensure alignment?
- Can we solve this by fixing the “context” in which AI is deployed, e.g. reforming human society?
- What kind of approaches might work for alignment? How can people contribute?
- How might alignment failures become apparent as AI progresses along the capability curve?

Add & vote on questions!

15:00-15:30 Scott Aaronson: Reform AI Alignment

(Slides) (Recording)

Abstract

Abstract:

I’ll share some thoughts about AI safety, shaped by a year’s leave at OpenAI to work on the intersection of AI safety and theoretical computer science.

Add & vote on questions!

15:30-16:00 Deger Turan: A Bold Attempt At Aligment: Open Agency Architecture

(Slides) (Recording)

Abstract & Background reading

- Abstract:

Open Agency Architecture is a bold theory and proposal for AI alignment that requires a massive and wide ranging formal-modeling enterprise that integrates into a global world-model. OAA systems do not deploy the trained ML system itself, but instead aim to constrain powerful ML systems to deploy verifiably aligned, less powerful outputs. Our plan is to develop OAA by iterating on smaller, domain-specific applications that can find immediate use as institutional decision-making tools and provide OAA with feedback from different academic disciplines and expert networks in an international collaboration.

- Background reading:

- The open agency model
- An open agency architecture for safe transformative AI
- Davidad's bold plan for alignment
- The open agency model
- An open agency architecture for safe transformative AI
- Davidad's bold plan for alignment

Add & vote on questions!

Chapter 3. Moral Characters of Crypto

This chapter aims to build a mental model of blockchains, cryptoeconomic mechanisms, and cryptography, focusing on their ability to align and coordinate agents.

16:00-16:20 Phil Daian: It's Too Late… MEV has Already Achieved AGI

(Slides) (Recording)

Abstract

Abstract

: In this talk, I will discuss the parallels between MEV alignment and AI alignment. First, I will give a brief introduction to MEV as a primitive for representing complex coordination games. I will argue that the MEV ecosystem represents a synthetic consciousness of fundamentally unaligned and often robotic actors, whose local incentives drive them to a common outcome. I posit several learnings and opportunities from the intersection of MEV and AI, including the ability to use cryptocurrencies as a hyper-realistic and ultra-adversarial sandbox to test agent modeling axioms. I will claim that privacy and decentralization are the key to an aligned future, and that we must align around these topics as humans as well.

Add & vote on questions!

16:20-17:00 Xinyuan Sun: Intelligence beyond Commitment Devices

(Slides) - (Recording)

Abstract & Background readings

- Abstract:

A major value proposition of cryptoeconomic mechanisms is that users can trustlessly collaborate by making credible commitments of their actions. We discuss ways where crypto-enforced credible commitments may mitigate human-AI coordination failures and demonstrate the limit and tradeoff of those commitment devices in mitigating intelligence alignment risks. We demonstrate how, surprisingly, the problem of mitigating the negative externalities of commitment devices in cryoto (i.e., MEV) is same as the problem of cooperative AI and a large part of AI alignment.

- Background readings:
- Foundations of cooperative AI
- Commitment games
- Program equilibria
- Reasoning about knowledge
- Ethereum is a game-changing technology
- Crypto as credible commitment devices
- Maximal Extractable Value (MEV) from commitment devices
- Speed of common knowledge in commitment devices | SUAVE
- Foundations of cooperative AI
- Commitment games

- Program equilibria
- Reasoning about knowledge
- Ethereum is a game-changing technology
- Crypto as credible commitment devices
- Maximal Extractable Value (MEV) from commitment devices
- Speed of common knowledge in commitment devices | SUAVE

Add & vote on questions!

17:00-17:15 Phil Daian Xinyuan Sun , Deger Turan: Open Agency Architecture Meets MEV: Collective Q&A

Add & vote on questions!

17:15-17:30 Barry Whitehat: Using Cryptography to prevent human defectors in world war AGI

(Slides) (Recording)

Abstract & Background Reading

- Abstract:

The crypto community has thought a lot about how to build collusion resistant mechanisms. Basically making it impossible to get bribed by making it impossible to prove you did the thing you want to get bribed for. If we combine this with proof of individuality and proof of possession of private key we can make it impossible for AI to bribe humans to defect.

- Background Reading:

https://vitalik.ca/general/2019/04/03/collusion.html

Minimal anti-collusion infrastructure - Applications - Ethereum Research

https://vitalik.ca/general/2019/10/01/story.html

Add & vote on questions!

17:30-17:45 Daniel Kang: What ZK can do for us: Privately Authenticating Real People (without third parties) and Auditing ML

(Slides) (Recording)

Abstract & Background Reading

- Abstract:

Zero-knowledge proofs have made amazing advances in proving arbitrary computation, but the real uses have mostly been limited in zkEVMs. In this talk, I will describe how to use zero-knowledge proofs to interact with the real world. I'll start by describing to authenticate real people and media (videos, images, audio) without

needing to trust third parties when combined with attested sensors. With open standards, we also don't need to rely on specific hardware vendors. I'll also describe how to audit ML deployments. As a case study, I'll describe how to audit the Twitter algorithm. The same technology can also be used to audit providers such as OpenAI.

- Background Reading

: * Verifying the Twitter algorithm

- Fighting deepfakes
- Using the zkml framework
- Verifying the Twitter algorithm
- Fighting deepfakes
- Using the zkml framework

Add & vote on questions!

17:45-18:00 Tea Break

6 Months Moratorium: Game Theory and Decision Theory

Chapter 4. The Art of Abstraction

18:00-18:30 Saffron Huang: Using the Veil of Ignorance to Align AI Systems with Principles of Justice

(Slides) (Recording)

Abstract

Abstract:

The philosopher John Rawls proposed the Veil of Ignorance (VoI) as a thought experiment to identify fair principles for governing a society. Here, we apply the VoI to an important governance domain: artificial intelligence (AI). In five incentive-compatible studies (N = 2, 508), including two preregistered protocols, participants choose principles to govern an Artificial Intelligence (AI) assistant from behind the veil: that is, without knowledge of their own relative position in the group. Compared to participants who have this information, we find a consistent preference for a principle that instructs the AI assistant to prioritize the worst-off. Neither risk attitudes nor political preferences adequately explain these choices. Instead, they appear to be driven by elevated concerns about fairness: Without prompting, participants who reason behind the VoI more frequently explain their choice in terms of fairness, compared to those in the Control condition. Moreover, we find initial support for the ability of the VoI to elicit more robust preferences: In the studies presented here, the VoI increases the likelihood of participants continuing to endorse their initial choice in a subsequent round where they know how they will be affected by the AI intervention and have a self-interested motivation to change their mind. These results emerge in both a descriptive and an immersive game. Our findings suggest that the VoI may be a suitable mechanism for selecting distributive principles to govern AI.

Add & vote on questions!

18:30-19:00 Tarun Chitra: Yudkowsky vs. Plato: can language models possess knowledge?

(Slides) (Recording)

Abstract & Background Reading

- Abstract:

Language models have ruined the ability for us to have a clean separation between man and machine — RIP Turing Test. On the other hand, other areas of computer science, such as interactive proofs and ZK have very ‘clean’ notions of knowledge built into their definitions. The type of ‘knowledge’ in zero knowledge exists in a particular sense — it can only ‘exist’ if a polynomial time algorithm generated it and it can only be ‘stolen’ if you have exponential compute resources. This dichotomy between the lack of “knowledge” in LLMs versus the formal and clear definition of “knowledge” in ZK suggests that we might be able to import some lessons about ‘knowledge’ from ZK to LLMs. In this talk, we’ll go through the epistemological concerns related to this question and try to provide some ideas for how LLMs can display possession of knowledge to each other.

- Background Reading:
- Initial Blog Post
- Chain of Thought Prompting Elicits Reasoning in Large Language Models (initial Google paper on CoT)
- Language Models (mostly) Know What They Know
- Justin Thaler’s ZK book, Section 7.4, Knowledge Soundness
- AI safety via debate
- Initial Blog Post
- Chain of Thought Prompting Elicits Reasoning in Large Language Models (initial Google paper on CoT)
- Language Models (mostly) Know What They Know
- Justin Thaler’s ZK book, Section 7.4, Knowledge Soundness

- AI safety via debate

Add & vote on questions!

19:00-19:30 Sarah Meyohas: Seaside Chat: can artificial intelligence produce art?

(Slides) (Recording)

Summary

Abstract:

In this talk, we will delve into the fascinating world of AI-generated images, exploring the intricate relationship between AI and artistic authorship. We'll examine whether searching for an image can truly be considered as creating it, and discuss the characteristics of a medium that influence the strength of authorship claims. Further, we'll investigate how the semantics of language as a tool for image synthesis impacts the end results and consider whether AI is inadvertently codifying "style" in its creations. Along the way, we'll ponder if AI-generated art evokes a sense of nostalgia by design, and ultimately, address the burning question: Can AI truly produce art?

Add & vote on questions!

Special thanks to Vitalik Buterin, Michael Johnson, Rob Knight, Nate Soares, Xinyuan Sun (Sxysun), Barry Whitehat, George Zhang and Zuzalu friends of the Pi-rate Ship.

Snacks for Thoughts…

Blockchains enable trustless collaboration via cryptographic and crypto-economic primitives. These primitives allow users to delegate their decision-making to smart contracts (algorithmic agents). And consensus on commitments makes this delegation common knowledge, thus shifting equilibria.

AIs, as complex algorithmic agents that may or may not employ agentic behavior, can lead to undesirable equilibria for humans, and many has even predicted to be bringing horrible destruction for humans very quickly. Can existing coordination technologies like crypto help us answer this question?

CryptoXAI delves into the coordination and alignment aspects of AI and crypto. After all, crypto's potential lies in its ability to act as a coordination device through the use of credible commitments, e.g., global payment, public goods funding, democratized financial access. How does crypto, as an alignment/commitment device, compare with popular alignment approaches such as decision theories or open-sourcing AI's source code? Does it make more sense to align AIs by combining functional decision theory with cryptographic commitments about the AI's actions instead of allowing arbitrary access to source code (which could cause programmable privacy issues)? But even if AIs can coordinate using cryptographic/crypto-economic commitments, can those commitments exercise be interpretable? What does the tradeoff space of those approaches look like?

Will AI use crypto to coordinate amongst themselves to improve the equilibria payoff? Will the equilibria that they coordinate align with the human social value? Can crypto as a commitment device be used to align AIs and humans? Afterall, some argues that AIs are still far from gaining agency and will stay in the"tool of humans" range for a long time. If that's the case, will the coordination and alignment of AI just ends up being a shadow of the coordination and alignment of humans. What unique properties would the projection of this shadow have? And if it does endup being human alignment, is it possible for crypto to exercise the coordination magic on humans to work on building AIs together (e.g., solving the data privacy training problem using some variant of orderflow auctions)? How about using crypto to coordinate humans in the period leading up to AGI, or to make the online world more secure against AGI?

What about AGIs? How fast will AIs gain agency? Does agency requireconsciousness? Does the development of AGIs lead to a world where there is one dominant Advanced AGI? Will AGIs have arbitrary preferences that make their alignment impossible? Will the access to privacy technology change what AGI could do (after all, the acess to source code would be impossible)?

What kind of human values can crypto, as commitment devices, align AIs to that wasn't possible with existing approaches like various decision theories? What are the limitations of commitment devices in its coordination of agents to reach human-valued outcomes? Can crypto learn from AI on how to best coordinate and trustlessly cooperate?