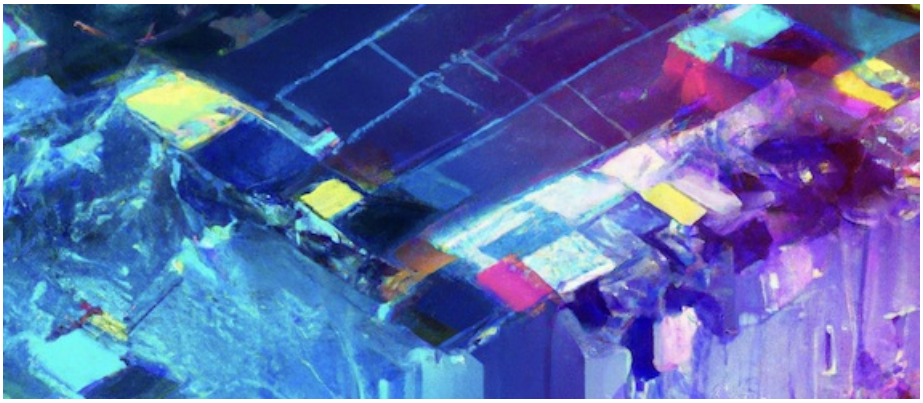
CredibleCommitments.WTF



Goals

After cryptoXai.wtf, we are moving the discussion around crypto and AI forward to a more focused direction - commitment devices. Specifically, we are having a small invite-only gathering/whiteboarding workshop.

In this event, we **focus** on answering two questions. One about commitment devices, another about AI.

1. **How we can better formalize and study the properties of crypto commitment devices and the coordination games mediated by it via the lens of commitment games?** Specifically, what kind of concrete simple games we can study to illustrate the power of crypto commitment devices.
2. **How can we better design blockchains (a technology for implementing commmitments) for coordination games with algorithmic agent participants (cooperative AI)?** What are some benchmark games that we can devise for boostrapping? e.g., are there any simple games where we can observe what kind of behavior emerges when AIs play games with the presence of a crypto commitment device.

This event is a series of open discussions on theWIP work of exploring practical implementations of cooperative AI. We specifically focus our discussion on concrete, approachable ways to study and realize commitments for cooperation. This event is to bring together facilitate discussions between pragmatic commitment device researchers and cooperative AI researchers to better map out the directions of this collaboration. There are two WIP motivation articles that give a flavor to the discussions:

1. **Why Cooperative AI and Crypto/MEV researchers should work together**
2. What can Crypto learn from AI

Agenda

Time: May 31th, 2023 - June 2nd, 2023

Location: Good Hotel @ London, colocated (~13 min walk) withAAMAS

Address: Good Hotel, Royal Victoria Dock, Western Gateway, London E16 1FA, United Kingdom

All time is UK time (BST)

Wednesday 1:00pm - 7:30pm Green Room (cabaret style)

Thursday 1:00pm - 7:30pm Pink Room (board room style)

Friday 8:30am - 1:30pm Pink Room (board room style)

Online Attandence

Zoom link: https://us06web.zoom.us/j/2466805965

Meeting ID: 246 680 5965

Find your local number: https://us06web.zoom.us/u/kcGTdTxYgg

Video Playlist

You can find a recording of the event atthis Youtube playlist.

People

Participants will be required to have high-context and have already read all materials in the reading list.

Participants include: DeepMind, Cooperative AI foundation, Ethereum Foundation, Flashbots, etc,.

Format

Small group whiteboarding and discussions (~10 people). There will be food catering and note takers. Discussion will be seeking technical depth.

Three day event.

One day for Permissionless Credible Commitment Devices.

One day for Cooperative AI.

One day to digest.

All of the sessions will be interactive whiteboard sessions. Free, relaxed environment.

Day 1: Blockchain Commitment Devices

A major value proposition of cryptoeconomic mechanisms is that users can trustlessly collaborate by making credible commitments of their actions.

Crypto-economic systems such as the Ethereum blockchain are commitment devices on which humans and algorithmic agents play real-world coordination games at large scale and high frequency. Importantly, the outcomes of these coordination games depend crucially on the design of the commitment device.

***In this day**, we aim to come up with a framework to study how the design of commitment devices change the values and incentives of the commitment games being played on top of the commitment device. In short, what is a good formalization for blockchain commitment devices that allows composition of knowledge?*

- *How we can better formalize and study the properties of crypto commitment devices and the coordination games mediated by it via the lens of commitment games?* Specifically, what kind of concrete simple games we can study to illustrate the power of crypto commitment devices.

13:00 - 14:00 Hangout

You can arrive early to meet and jam with other participants.

14:00 - 15:00 Illustrations on Alignment and Commitment in Decentralized Systems

led by David Parkes. slides. video

It will be an overview talk, and then open up for Q&A discussions.

Description

- Differentiable economics for contract learning
- Verifiable sequencing rules for MEV, as an example of policing the commons
- Decentralized, faithful implementations
- Strategyproof computing

15:00 - 16:00 Aligning like a protocol

led by Barnabe Monnot. slides. video

Description

The Ethereum protocol embodies the will of the Ethereum community, which sets up the protocol to automate the provision of blockspace by a decentralised set of operators. The protocol specifies incentives for these operators, expected to align

their behaviour with the will of the community. Yet the protocol does not “see” all things that validators do, and some functions, such as ordering transactions in a block, are left unspecified. This opens the door for “metagame” which augment the action set and shift the payoffs of operators beyond what the protocol specified for them.

The question of aligning the protocol-as-will-of-the-community and its operators,**is similar to the framework developed by the Cooperative AI research programme**.

- By building in more Introspection/Understanding, the protocol becomes “aware” of the games being played, and is better able to control its outcomes
- By extending its Bridges/Communication with further domains, it gathers intelligence which may help it achieve higher welfare outcomes
- Exercising its Agency/embodying a Commitment to a course of action, the protocol is able to deter certain attacks, and provide credibility for its mechanisms as a public good
- Finally, the protocol aims to become an Institution, complete with a loose governance process, which strengthens in time the community-determined outcomes that it aims to achieve.

16:00 - 17:00 Mediator Extractable Value on Permissionless Credible Commitment Devices

led by Xinyuan (Xyn) Sun. slides. video

Description

We discuss ways to think about permissionless credible commitment devices through the lens of cooperative game theory and commitment games, with the (un)surprising observation that Maximal Extractable Value (MEV) arise as a necessity from the design of such devices: due to the specific design of blockchain as a commitment device (blocktime, etc), the mediators of blockchain commitment devices is able to extract value from the games that is being played on top of the commitment device, where the value extraction process causes huge negative externalities.

In this session, we first goes through an intro to crypto and MEV. And then brainstorm the existing open problems in studying crypto and MEV via commitments. Afterall, as soon as we try to make an impact and actually implement any commitment system (for AIs or for humans), we will run into problems that involve design choices of what substrate those commitment games are played on, and those design choices entail the study of MEV.

Readings

- SUAVE via commitments

17:00 - 18:00 Routing MEV in Constant Function Market Makers

led by Kshitij Kulkarni. video

Description

Constant function market makers (CFMMs) are a subclass of decentralized exchanges that have been used to trade trillions of value over the last few years. These exchanges are susceptible to a particular kind of extractable value by validators and searchers known as sandwich attacks. We analyze sandwich attacks in a setting in which users want to route trades over a network of CFMMs. We define two notions of routing: selfish routing, in which users route to maximize their pro-rata share of the output from each path, and optimal routing, in which a central planner can coordinate to route the trades to maximize the output from the network. We then introduce sandwich attacks in this setting, and show the existence of an "inverse Braess paradox": sandwich attackers can in-fact improve the price of anarchy of this system. We further construct general price of anarchy bounds.

Readings

- Credible, Optimal Auctions via Blockchains

18:00 - 19:00 Fireside Chat: Anthropic Alignment

led by Georgios Piliouras and Vincent Conitzer. video

This is a panel. Tentative question list 1 and list 2.

19:00 Commitment Devices dinner

Dinner to continue discussion and for chill hangouts.

Day 2: Cooperative AI

AI coordination and alignment are necessary. Traditional social technologies like policies and regulation are inadequate for coordinating games with AI agents.

Crypto-economic commitments provide an important technology for coordination and alignment of AI agents compared to traditional social technologies. For example, the crypto sandbox offers a real-world environment for experimenting with AI coordination, while the study of commitments in crypto provides a pragmatic framework for understanding commitment devices and their limitations.

***In this day**, we explore the synergy between cooperative AI and cryptoeconomic mechanisms via bridge of credible commitments. We aim to find concrete examples where blockchain researchers and AI researchers can work together. For example, how crypto-economic commitments actually offers the medium through which agents can verify the program that each other is running on (as in program equilibria). We will compound the learnings from day one and try to answer the following:*

- *How can we better design blockchains (a technology for implementing commmitments) for coordination games with algorithmic agent participants (cooperative AI)?* What are some benchmark games that we can devise for boostrapping? e.g., are there any simple games where we can observe what kind of behavior emerges when AIs play games with the presence of a crypto commitment device.

13:00 - 14:00 Hangout

You can arrive early to meet and jam with other participants.

Lunch will be delivered between 1-1:30pm from The Real Greek and a separate Uber Eats order will come with beverages

14:00 - 14:15 Why Cooperative AI and Blockchain Researchers Should Collaborate

This will be an interactive session where we lay out the foundation of the correspondences between two fields.

14:15 - 15:15 Cooperative AI in the Real World

led by Lewis Hammond. video

Description

In this session I will provide a brief introduction to the topic of cooperative AI, before diving into a discussion of where it might be applied in the real world. Thus far, there have been relatively few settings in which sophisticated autonomous agents engage in mixed-motive interactions with one another, and with humans. In the coming years, this looks set to change. If we want to drive progress on improving the cooperative capabilities of advanced AI systems, it will be important to identify real-world domains that they can be safely tested in. I will sketch some desiderata and possibilities for these domains, aiming to solicit suggestions and disputations from attendees.

15:15 - 15:45 Formal Contracting for Multi-Agent Systems

led by Andreas Haupt. video

Description

We draw upon the idea of formal contracting from economics to overcome diverging incentives between agents in multi-agent systems. We propose an augmentation to a Markov game where agents voluntarily agree to binding state-dependent transfers of reward, under pre-specified conditions. Our contributions are theoretical and empirical. First, we show that this augmentation makes all subgame-perfect equilibria of all fully observed Markov games exhibit socially optimal behavior, given a sufficiently rich space of contracts. Next, we complement our game-theoretic analysis by showing that state-of-the-art RL algorithms learn socially optimal policies given our augmentation. Our experiments include classic static dilemmas like Stag Hunt, Prisoner's Dilemma and a public goods game, as well as dynamic interactions that simulate traffic, pollution management and common pool resource management. I discuss differences of agentic and technological augmentations, delegation versus transfers, limitations to contractible outcomes and its relationship to moral hazard, and coordination issues in commitment.

Readings

- Moral Hazard and Observability
- Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL

16:00 - 17:00 Surrogate goals and safe Pareto improvements

led by Caspar Oesterheld. video

Description

A set of players delegate playing a game to a set of representatives, one for each player. We imagine that each player trusts their respective representative's strategic abilities. Thus, we might imagine that per default, the original players would simply instruct the representatives to play the original game as best as they can. In this paper, we ask: are there safe Pareto improvements on this default way of giving instructions? That is, we imagine that the original players can coordinate to tell their representatives to only consider some subset of the available strategies and to assign utilities to outcomes differently than the original players. Then can the original players do this in such a way that the payoff is guaranteed to be weakly higher than under the default instructions for all the original players? In particular, can they Pareto-improve without probabilistic assumptions about how the representatives play games? In this talk, I'll show that the answer is yes!

Readings

- Safe Pareto Improvements for Delegated Game Playing

17:00 - 18:00 Private information and miscoordination: key challenges for Cooperative AI

led by Anthony DiGiovanni. video

Description

As has been shown in the literature on program equilibrium, credible conditional commitments can allow for mutual cooperation (in the sense of Pareto efficiency) to be rational in social dilemmas, such as the Prisoner's Dilemma. Even when rational agents have access to credible commitment mechanisms, however, they may fail to cooperate for two main reasons. First, they may have private information that they do not have the ability or incentive to disclose. Second, they may fail to coordinate on one cooperative equilibrium, when they have different preferences over different efficient outcomes. We will discuss (1) how certain kinds of conditional commitments can in theory mitigate both these causes of cooperation failure, when agents have access to some kinds of verification technology (in part inspired by Safe Pareto Improvements); and (2) how these theoretical ideals might be practically implemented with techniques from cryptography.

Readings

- Commitment Games with Conditional Information Disclosure
- Improved coordination with failsafes and belief-conditioned programs

18:20 - 19:30 Benchmark/contest for cooperative AI with pragmatic commitment devices.

led by Lewis Hammond and Xinyuan Sun

Description

Can do mindmapping. This will be a brainstorming session where we think of the concrete problems that crypto-economic commitment devices could implement for cooperative AI. Some sample questions:

- how AIs will behave in the presence of commitment devices <> how will AI deploy commitments and use those as a part of their action space
- what are sensible AI applications to deploy on top of crypto commitment devices.
- recommender systems and ads that they show
- Cicero but AI can read Ethereum validator code and can run the Ethereum code, how will they behave?
- colored agents and their behavior change in presence of commitments
- Prompt Injection
    - proprietary data usage
    - security against targeted adversarial activity
    - modularity of datasets & prompts; efficient incremental finetuning

19:30 Cooperative AI via Crypto-economic Commitments Dinner

Dinner to continue discussion and for chill hangouts.

8pm at Chai Ki, Canary Wharf under name "Laura"

Day 3: Digestion

***In this day**, we connect our learnings from the previous discussions.*

8:30 - 9:30 Hangout

You can arrive early to meet and jam with other participants.

9:30 - 10:30 What did we learn? Sharing takeaways.

10:45 - 12:00 Knowing what we know from the workshop session, where do we go now and in the future?

Reading List

Crypto

- Blog: Ethereum is game-changing technology, literally.
- Game Mining: How to Make Money from those about to Play a Game
- Stackelberg Attacks on Auctions and Blockchain Transaction Fee Mechanisms
- Game theory on the blockchain: a model for games with smart contracts
- Blockchain Mediated Persuasion
- Credible, Optimal Auctions via Blockchains
- Costs of Sybils, Credible Commitments, and False-Name Proof Mechanisms
- Optimal Routing for Constant Function Market Makers
- Towards a Theory of Maximal Extractable Value I: Constant Function Market Makers
- Credible Decentralized Exchange Design via Verifiable Sequencing Rules
- Slides: MEV and Credible Commitment Devices
- Slides: Intelligence Beyond Commitment Devices
- Video: An Overview of Credible Commitment Devices
- Blog: Mutating Mempools and Programmable Transactions
- Video: Smart Transactions
- Blog: SUAVE through the lens of game theory

Credible Commitments

- Commitment games
- Program Equilibrium
- Strong Mediated Equilibrium
- A Commitment Folk Theorem
- A folk theorem for Bayesian games with commitment
- Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory
- A Note on the Compatibility of Different Robust Program Equilibria of the Prisoner's Dilemma
- Commitment Games with Conditional Information Disclosure
- Strategyproof Computing: Systems Infrastructures for Self-Interested Parties
- Mechanism Design With Limited Commitment
- Computing the Optimal Strategy to Commit to

- Improved coordination with failsafes and belief-conditioned programs

Cooperative AI

- Course: Foundations of Cooperative AI
- Foundations of Cooperative AI
- Open Problems in Cooperative AI

Misc

- Blog: Why Cryptoeconomics and X-Risk Researchers Should Listen to Each Other More
    - AI safety is about agents with IQ 150 trying to control agents with IQ 6000, whereas cryptoeconomics is about agents with IQ 5 trying to control agents with IQ 150
- Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection