Why Cooperative AI and Blockchain Researchers Should Collaborate

TL;DR: AI coordination and alignment are *necessary*. Traditional social technologies like policies and regulation are *inadequate for coordinating games with AI agents*. Crypto-economic commitments can provide a key technology for the cooperation among AI agents and help to overcome some of the shortcomings of traditional social technologies. **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.**

After making the above arguments, we illustrate the synergy between *crypto infrastructure* and *AI agents* via a few relevant examples:

1. Congestion Games
2. Credibility of Commitments
3. Limits of Real-world Commitments
4. Competing Commitments
5. Conditional Commitments
6. Parallels with Decision Theories

Background

Crypto-economic systems such as the Ethereum blockchain are commitment devices on which *humans and bots play real-world coordination games* at large scale and high frequency.

The outcomes of these coordination games depend crucially on the design of the commitment device. The area that studies how the design of commitment devices change the values and incentives of the commitment games being played on top of the blockchain is called *Maximal Extractable Value* (MEV).

For more information, readers can refer to the References at the end of this document.

AI Coordination is Necessary

The current state of AI, coupled with its future potential for increased sophistication and possible misuse in an unrestricted action space, endows it with the ability to cause great harm.

For example, recommender systems in the attention economy have already impacted large scale political campaigns, created an increasingly fractionalized society where the *common knowledge of contextual semantics is lost* This crisis is exacerbated by the generation and dissemitation of misinformation where humans are *"eclipse attacked" by AIs with siloed information echo chambers*.

This capability to mislead humans poses a substantial threat to vital societal structures that we rely on for coordination such as trust, corporations, and governments. Why? Because common knowledge, which is the base of coordination, is lost.

Moreover, huge negative externalities often arise from the unrestricted empowerment of algorithmic agents, leading to far-reaching effects due to their lack of coordinated interaction. Examples include:

1. **Traditional financial system**
    - The infamous flash crash events are a prime example. Here, algorithmic trading bots, acting independently, induce economic turbulence by amplifying market volatility. This leads to dramatic price fluctuations that bear no resemblance to the underlying asset's true value.
2. **Online marketplaces**
    - We also witness these effects in eBay auctions. Sniper bots, operating without coordination, drive the price of commonplace goods to ludicrous levels due to uncontrolled recursive behavior.
    - Market health suffers when individually trained models, acting without coordination, engage in collusive pricing. In this scenario, the models set prices that disadvantage the majority of market participants.
3. **Onchain economies**

- Adversarial Player vs. Player (PvP) "bot wars" bring about substantial negative externalities for regular users of the commitment device. These uncoordinated activities exacerbate congestion fees, causing ordinary users to pay more than necessary. The credibility of the system is also undermined through time-bandit attacks and reorgs. Furthermore, the very existence of the commitment device is threatened due to the centralization of the validator set and vertical integration, reducing the pool of disinterested parties maintaining the commitment device (i.e., the mediator/auctioneers collude or deviate).

Inadequacies of Traditional Coordination Tools for Cooperative AI

In face of this crisis, our traditional tools for coordination and alignment, like regulation or policies, are not sufficient for the task of coordinating games where participants include algorithmic agents. Moreover, they often impede innovation by restricting growth and experiments, and entrench centralized players who has political capital or is in a previledged position to play power dynamics.

For example, despite numerous attempts, governments and regulators worldwide have struggled to effectively control tech giants, evidenced by challenges in areas such as data privacy (Facebook's Cambridge Analytica scandal), antitrust enforcement (Big Tech dominance), content moderation (misinformation on social media), encryption dilemmas (conflicts with law enforcement), labor laws (gig economy issues), and AI regulation (ethical use of AI).

The **complexity** and **rapid evolution** of these technologies, coupled with their **global reach**, further exacerbate these regulatory hurdles. The instability of US politics will make the regulation landscape and policy for AI coordination especially hard for the years ahead: political polarization and frequent changes in administrations often cause regulatory lags. But we need to make progress with coordination games of algorithmic agents ASAP, not only for bounding the worst case (safety) but also in enabling the best case (efficiency).

The Choice is Obvious

The ability to form binding, credible commitments efficiently and flexibly offers a wealth of of possibilities for the development of more cooperative AI systems. It forms an important complement to other abilities, such as improved communication, better modelling of other agents, and the creation of norms and for AI systems. These technical approaches in turn form a complement to the traditional social technologies of policy and regulation, which, as argued above, are insufficient for ensuring coordination in the context of sophisticated AI agents.

Without an effort to imbue our AI agents with improved cooperative capabilities, such as crypto-economic commitment devices, we risk facing crises and/or power imbalances at an unprecedented scale. We therefore advocate for the development of an AI-compatible crypto-economic infrastructure of permissionless credible commitment devices.

The Inevitable Evolution of Market Structures

The majority of coordination games currently being played are slow/local games (e.g., a group of oil drillers deciding on their respective share). However, with the rise of AI algorithmic agents, an increasing number of games will be played at a high-frequency setting on a global scale, i.e., they will be fast/global games (e.g., two bots bidding in some auction). We can anticipate an ultra-refined service/labor market where ad-hoc agreements can be made from time to time with various parties who are better suited for the task at hand, eliminating the need for long-term commitments with familiar parties.

These games predominantly occur in high-frequency, low-latency environments, necessitating a coordination technology capable of handling rapid and global dynamics.

Traditional methods such as trust or violence enforcements are ill-equipped to meet these demands due to their inherently slower monitoring speed and the added burden of extensive monitoring that entrench centralized parties. In other words, the coordination technology we had before is tailored for slow and local games, not for **fast and global games**.

As we advance in the field of AI, we foresee an era of **ubiquitous high-frequency trading (HFT) and hyper-financialization**. AI, in its quest for efficiency, will lead to the creation of markets for nearly every conceivable service or commodity. The coordination games associated with these markets are subject to frequent structural changes, making it infeasible and inefficient to coordinate them by imposing restrictive structures. As demand and supply interact in real-time, new games are constantly being generated.

In this dynamic landscape, mechanisms are transient, and millions of bots operate momentarily to achieve a Turing-complete market. Such games are optimally coordinated via commitments that shape agents' incentives on an ad-hoc basis. These commitments can be effectively managed with low verification costs, facilitated by low-latency cryptography and programmable privacy, providing a solution that is as agile as it is efficient.

Moreover, studying the coordination of those fast and global games (e.g., searchers bots backrunning each other) is where we should start, because they are more technical and therefore easier to scrutinize due to their relatively certain structure (uncertainty tends to increase with game duration).

We have a reasonable understanding of how to coordinate fast games because they can be structured, allowing us to apply

our knowledge from mechanism design and game theory. In contrast, slow games carry so much uncertainty that it often proves counterproductive or even harmful to attempt to structure them. This is why the coordination of slow games frequently devolves into power struggles and politics, which makes it hard to make concrete progress on the AI coordination problem. **For these reasons, crypto presents a tractable path forward for AI coordination and alignment.**

We should start experiments in AI coordination with crypto-economic games because the game is far more manageable and certain - and we know how to solve it. After accumulating sufficient data or experience on how AIs behave or coordinate in these fast games, we can progress to extend the game's duration. The argument here is that slow games bear more resemblance to single-agent alignment due to the uncertain game structure. It's always possible for someone to discover an uncoordinated game with a payoff correlated to your existing coordinated game, causing the existing incentive Rube Goldberg machine to collapse. Fast games, however, are more akin to multi-agent alignment because their shorter duration makes them more certain.

In essence, when coordinating a slow game, you aim to engineer player incentives in the right direction robustly, to new dynamics and new games being played. This is often impractical to do programmatically, explaining why slow game commitments (e.g., laws, policies) often retain a human discretion element. These commitments need to confront unknown unknowns or fundamental uncertainties, which can only be effectively handled by human intelligence.

Alternatively, the commitment can choose to sacrifice efficiency for robustness, meaning it deliberately reduces game structure uncertainty via constraining the action space. But for fast games, the action space is usually limited, so algorithmic agents (commitments) don't have to deal with much uncertainty. As such, there's significantly less demand for increasing the robustness of the commitments. Furthermore, these commitments don't need to be complex because they are expected to expire shortly after the fast game concludes.

The Dual Value Prop

There are two major value adds of crypto (and especially the field of MEV) to AI:

1. a **pragmatic framework for understanding commitments** and the intricacies of commitment devices
2. a **rigorous real-world sandbox with incentivized and hyper-adversarial agents**, with automated symbolic behavior on a massive scale, that is perfect for testing the limit of AI coordination.

Prop 1: Pragmatic Framework for Programmatic Commitments

Crypto is familiar with the study of commitments and how various designs of different commitment devices could impose unexpected outcomes to the games being coordinated on top of those devices, both from the practical side and the theoretical side.

For example, due to the specific design of blockchain as a commitment device (blocktime, fee schedule, inter-temporal monopoly by commitment device mediator, credibility coming from anonymous set of validators and sometimes centralized offchain systems that are blackboxes), the mediators of blockchain commitment devices is able to extract value from the games that is being played on top of the commitment device. And this value extraction process causes huge negative externalities. The study and mitigation of such negative externalities caused by different commitment device design is MEV, which focuses on the study of commitments in adversarial environments with high-frequency games.

Ultimately, as long as we are trying to make an impact and actually implement any AI commitment system, 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.

Prop 2: The Adversarial Sandbox

Crypto has already implemented usable and battle-tested commitment devices where there exists many adversarial attackers who operate beyond the scope of the commitment device that interact with well defined agents. It is exactly akin to the problem of ["how can a dumb reasoner protect itself against smart intelligences"](#) where the "dumb reasoner" are transactions that are not expressive (i.e., a shallow copy of the user's intent), and yet it has to interact with MEV searcher bots who are way more "intelligent." And in crypto, the way we mitigate the adversarial environment is via conditional commitments (i.e., the weaker party only interacts conditional on a voluntary restriction of the stronger party).

Moreover, the study of those commitment devices in crypto is already about algorithmic agents that look like AIs that we care about: **many activities on blockchain are driven by trading firms whose entire goal is to optimize a scalar, and their systems are largely uninterpretable**, and yet we can successfully coordinate their behavior to some extent (this is contrary to many AI alignment belief that one cannot align intelligences that one don't understand, we don't understand humans and yet we've achieved many coordinated outcomes).

Such adversarial behavior against commitment and algorithmic agent coordination games on the blockchain are very prevalent and at scale (current lower bound estimate of a billion of USD per year), so there are enough incentive and space to test the robustness of existing alignment and cooperative AI studies.

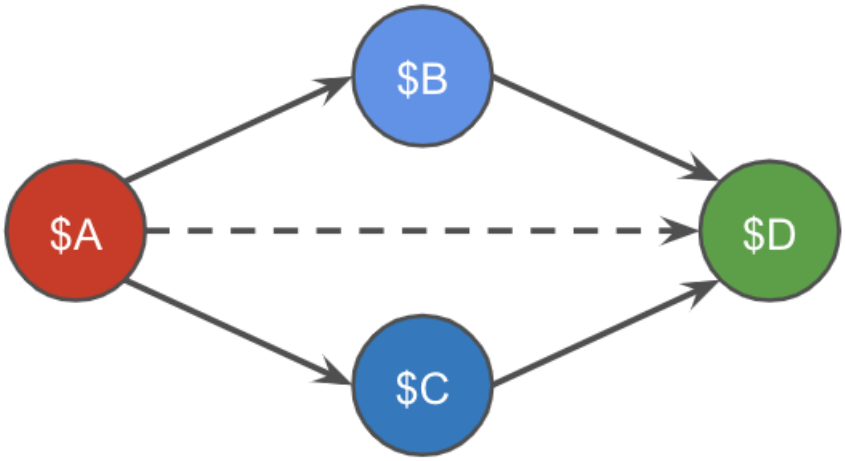
Examples

Example I: Congestion Games

Congestion games have been a major area of interest for computational game theory and multi-agent systems for the past two decades. They are a canonical example of a mixed-interest game in which there are both competitive and cooperative motives at play. Recent studies in artificial intelligence have focused on algorithms for learning equilibria and welfare effects, in different subclasses of games that make different assumptions on the information available to agents.

Despite the compelling example and theoretical interest, currently we lack a concrete environment in which to test AI agents playing such games. Real world AI agents such as self driving cars are only starting to be slowly and selectively deployed, and they constitute a small fraction of the players in a routing game dominated by human drivers.

We argue that blockchains offer a unique environment to test how self-interested agents behave in a special but representative class of congestion games: financial orders routing. In particular, the Decentralized Finance (DeFi) infrastructure on Ethereum offers an ideal environment for such tests where: (1) users submit orders in real-time 24/7, (2) costs/incentives are real and substantial, (3) all agents routing orders across the financial network of exchanges/counterparties are bots, (4) the system is a permissionless API so new AI bots can be tested, (5) data is mostly public so it can be used for analysis, training, and simulation.



Cartoon DeFi network with 4 assets and routes between them. (Full lines are smart contract exchanges and dotted line is a (potential) direct swap.)

Moreover, such an environment has: a rich evolving topology with new nodes (assets) and new routes (liquidity pools, exchanges, blockchain bridges) being added, a differential set of more and less sophisticated/informed agents that create and route orders. On the infrastructure side, there is a lot of research at the intersection of MEV and routing, see cross-domain MEV or routing games with MEV extraction. Complex behaviors have emerged such as sophisticated strategies for frontrunning or backrunning orders that are routed suboptimally, from an ecosystem of self-interested agents and organizations. The blockchain environment also allows to deploy new mechanisms/commitment devices through which more sophisticated orders can be expressed to improve coordination.

Example II. Conditional Commitments

The crypto infrastructure really shines in its ability to implement, via smart contracts, practical commitment devices that are neutral, programmable, and credible. Conditional commitments have been widely studied at the intersection of game theory and cooperative AI (commitment games and program equilibria). In blockchain, there are several projects that are implementing intent-centric architectures that have conditional commitments as primitives, to enable a set of principals to express rich preferences that allow them to coordinate. For example, Flashbot's SUAVE and Anoma are two projects that are tackling the engineering challenges with building such systems. Let me be more explicit, **intents are programs and these cryptoeconomic protocols implement the program equilibria studied in AI**. Moreover, these protocols have cryptographic gadgets that allow to handle the incomplete/asymmetric information cases as well.

As a concrete example we use the DeFi order routing game discussed above to illustrate how these intents work. A user intent could be the following

```

if exists swap $D -> $A at rate $A/$D <= 1000:

swap 500 $A -> $D
```

else:

do nothing

```

This is a simple intent that checks if there is an intent from a user willing to swap in the reverse direction at acceptable price, which creates an additional direct route `$A -> $D` in the DeFi network, in which case it will resolve in a pair of mutual swap actions, otherwise the intent does nothing or reverts. Note that another option could be to specify that, in the `else` case when there is no matching swap, the order could be routed along the most efficient exchange route between `$A -> $B -> $D` and `$A -> $C -> $D`.

We can easily immagine scenarios where this architecture unlocks more efficiency than the "blind orders" that are used today where each user sends a transaction to the network to swap on a fixed route, without conditioning on what others are simultaneously doing. Another example which is slightly more sophisticated is for the principal/user to submit intents that allow routers/agents to coordinate on pareto-superior equilibria, for example with a program that encodes the following: *I commit to change route if the other agents commit to transferring me a fraction of the surplus they will get as a result of my action*.

## Example III. Decision Theories

Crypto commitment devices also offer interesting thought frameworks for thinking about decision theory problems in cooperative AI. For example, flashloans in crypto are commitments that have a decision/action in the logical-time past depending on a decision/action in the logical-time future, this is possible because of the atomicity of transaction semantics on blockchains.

Specifically, one could say: "I commit to lending you all my money if there exists a commitment in future from you that returns all my money." Smart transactions is a generalization of flashloans that expand the semantics of blockchain transactions to native support such "time-travelling" behavior natively by allowing interdependent commitments at a larger scale. Basically, flashloans are a way to generalize "causality" in the distributed systems logical time sense to "correlation."

One could observe a striking similarity between flashloans and classical problems in decision theory, such as newcomb's paradox, where Omega is essentially instantiating a structure like a flashloan when making the prediction of user's choice (Omega's decision in the past depends on the user's choice in the future). And the way blockchain transaction semantics are defined corresponds with many ideas in evidental decision theory or functional decision theory, where "causality" is weakened and "correlation" (or, commitments that depend on commitments that exists in "future" logical time) is emphasized in resolution and settlement of commitments.

## Example IV. Competing Commitments

Another interesting aspect of commitment devices is how they live in the presence of competing commitment devices. For example, one strategy that was observed from algorithmic trading bots is that they "commit" to not being able to commit to anything, which therefore makes other agents' commitments (credible threats) obsolete. Another example is that one commitment device can attack another commitment device by providing a commitment that releases a bounty if the other device breaks the credibility of previously settled commitments (i.e., reorg-as-a-service).

## Example V. Limits of Real-world Commitments

Another example is that implementing commitments takes time and resources (e.g., implementing common knowledge via consensus protocols takes significant time if we seek geographical decentralization and credibility of finalized commitments), and this time might be too long to coordinate some games (e.g., financial markets move 24-7 and the game played there is continuous and way faster than the commitment implementation time), and the lack of coordination of those games could cause ripple effects in the coordination of other games (e.g., the market marker structure may need to change to accomodate for the fact that fast games cannot be coordinated, like giving unsophisticated liquidity providers more fees, or that sophisticated agents could frontrun or sandwich unsophisticated agent's commitments). Or, if the commitment device gives mediators the power of inter-temporal monopoly, then the mediator is able to extract the full surplus from coordination games that are played within the commitment credibility window, making the agents indifferent between using the device or not.

## Example VI: Credibility of Commitments

Another example is that many cooperative AI or AI alignment literature doesn't justify the assumption of the commitment credibility, while in reality if we implement those coordination techniques, the design choices of the commitment device vastly impacts the credibility of the commitment and how the commitment games are played out (e.g., agents exhibit vastly different behavior when the commitment devices are designed differently, and they may cause second-order coordination difficulties).

Aside from not directly addressing commitment designs, in cooperative AI, it is often directly assumed that the commitments that AI make will be credible and the settlement of those commitments are perfect and robust with respect to other
```

competing commitments. Also, AI literature tends to assume AIs open-sourcing their own source code as a form of credible commitment, this is basically saying the AI model needs to be committed on-chain, because otherwise in an adversarial asynchronous network environment, just broadcasting one's source code is not credible as for example other AIs could simply pretend to have never received the package. All of those assumptions needs to be justified. And crypto provides the place for studying the justification of those and study how would such AI coordination proposals stand the test of adversarial attackers.

For example, if the commitment device orders commitments in a first-come-first-serve way, then agents have an incentive to vertically integrate with the mediators, which causes the credibility of the commitments to go down and loses its initial purpose of pursuading the other agents that the commitment is subgame-perfect and therefore can shift equilibria. Moreover, in order to keep the commitments credible, there needs to be an incentive for the mediators (or else running the commitment device becomes a public good) that is roughly poportionate to the amount of economic value that is generated from the commitment games that is being played on top. And the economics (both supply, how do we charge fees for those incentives, and demand, how do we distribute the collected profits to mediators who provide differentiated services) of mediator incentives turn out to be non-trivial, and, if not well-designed, causes huge negative externality and can completely misalign the commitment games played on top.

Why Blockchain?

Previously we argued that the only technology that is suitable for implementing AI coordination is using blockchains. But why blockchains? Why can't a consortium of AI labs just build a non-crypto protocol that operates a semi-permissionless commitment device for coordination games with algorithmic agents?

Afterall, in this way, the commitment device design can be updated agilely as there is less social coordination cost, and that the system can be more efficient because now only entities that are relevant to playing the game have to provide the mediation service. For example, the commitment device will have lower fees and fewer MEV, and it will still be able to coordinate high-frequency fast/global games, and is free to use as long as people trust the consortium.

The reason is because those systems are not robust.

If the commitment device is not truly decentralized, then it is hard for new entrants to compete with existing players. This is bad not only because the existing players could extract rent (and sometimes this rent is undiscoverable), but more importantly, those existing players are entrenched despite they don't have the most up-to-date mechanism that adapts to the rapid evoluting market structures. For example, in such a consortium system, new entrants have to first play political games to gain trust from the consortium of AI labs before they can enter the market to compete, and political games are hugely centralizing in that they encourage nepotism and entrench previledged players. So even if a new market participant has a better mechanism in mind, she cannot compete with existing players because the acquisition of trust, a centralized resource, is hard and a gatekeeping process.

Ultimately, high welfare can only be achieved if we have decentralization in the sense ofHayekian “use of knowledge in society.”

In the face of fundamental uncertainty in the world, no centralized consortiums can possibly have confidence in “the mechanism we have at hand is the best mechanism for implementing AI coordination and commitment games and is robust in face of future changes,” especially when the market structure is constantly changing and each morphism could possibly break the alignment and coordination of existing games because they bring in uncoordinated new games that have correlated interests with existing games. Therefore we must design the market in a way such that cooperative AI mechanism designers can and are incentivized to freely compete.

Robustness in cooperative AI solutions can only be achieved via free market competition in the commitment mechanism market. And free market competition can only be achieved via decentralization.

Conclusion

In conclusion, crypto is the only way to implement cooperative AI. Crypto commitments provide a superior technology for coordination and alignment of AI agents compared to traditional social technologies like policies and regulation. The crypto sandbox offers a rigorous, real-world environment for testing the limits of AI coordination, while the study of commitments in crypto provides a pragmatic framework for understanding commitment devices and their limitations. As we advance in the field of AI, it is important to recognize the role that crypto commitments can play in addressing the challenges of coordination and alignment.

References

Blockchain commitment devices (also called *permissionless credible commitment devices* or PCCDs)

- 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
- Video: Designing Smart Contracts With Free Will

MEV

- Flash Boys 2.0: Frontrunning, Transaction Reordering, and Consensus Instability in Decentralized Exchanges
- Clockwork Finance: Automated Analysis of Economic Security in Smart Contracts
- Towards a Theory of Maximal Extractable Value I: Constant Function Market Makers
- Optimal Routing for Constant Function Market Makers
- Price of MEV: Towards a Game Theoretical Approach to MEV

Misc

- Slides: MEV and Credible Commitment Devices
- Slides: Intelligence Beyond Commitment Devices
- Video: An Overview of Permissionles Credible Commitment Devices
- Blog: Mutating Mempools and Programmable Transactions
- Video: Smart Transactions
- Blog: SUAVE through the lens of game theory