Why Cryptoeconomics and X-Risk Researchers Should Listen to Each Other More

Vitalik Buterin

Follow

--

27

Listen

Share

Special thanks to Jaan Tallinn for early feedback and comments.

There has recently been a small but growing number of signs of interest in blockchains and cryptoeconomic systems from a community that has traditionally associated itself with artificial intelligence and various forms of futuristic existential risk research. Ralph Merkle, inventor of the now famous cryptographic technology which underpins Ethereum's light client protocol, has expressed interest in DAO governance. Skype co-founder Jaan Tallinnproposed researching blockchain technology as a way to create mechanisms to solve global coordination problems. Prediction market advocates, who have long understood the potential of prediction markets as governance mechanisms, are now looking at Augur. Is there anything interesting here? Is this simply a situation of computer geeks who were previously attracted to computer-geek-friendly topic A now also being attracted to a completely unrelated but also computer-geek-friendly topic B, or is there an actual connection?

I would argue that there is, and the connection is as follows. 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?

In the context of AI research, a major sub-problem is that of defining a utility function that would guide the behavior of a superintelligent agent without accidentally guiding it into doing something that satisfies the function as written but does not satisfy the intent (sometimes called "edge instantiation"). For example, if you tried to tell a super-intelligent AI to cure cancer, it may end up reasoning that the most reliable way to do that is to simply kill everyone first. If you tried to plug that hole, it may decide to simply permanently cryogenically freeze all humans without killing them. And so forth. In the context of Ralph Merkle's DAO democracy, the problem is that of determining an objective function that is correlated with social and technological progress and generally things that people want, is anti-correlated with existential risks, and is easily measurable enough that its measurement would not itself become a source of political battles.

Meanwhile, in the context of cryptoeconomics, the problems are surprisingly similar. The core problem of consensus asks how to incentivize validators to continue supporting and growing a coherent history using a simple algorithm that is set in stone, when the validators themselves are highly complex economic agents that are free to interact in arbitrary ways. The issue found with the DAO was a divergence of software developers' complex intent, having a specific use in mind for the splitting function, and the de-facto result of the software implementation. Augur tries to extend the consensus problem to real-world facts. Maker is trying to create a decentralized governance algorithm for a platform that intends to provide an asset with the decentralization of cryptocurrency and the reliability of fiat. In all of these cases, the algorithms are dumb, and yet the agents that they have to control are quite smart. 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 — problems that are certainly different, but where the similarities are not to be scoffed at.

These are all hard problems, and they are problems that both communities have already been separately considering for many years and have in some cases amassed considerable insights about. They are also problems where heuristic partial solutions and mitigation strategies are already starting to be discovered. In the case of DAOs, some developers are moving toward a hybrid approach that has a set of curators with some control over the DAO's assets, but assigns those curators only limited powers that are by themselves enough to rescue a DAO from an attack, but not enough to unilaterally carry out an attack themselves that causes more than moderate disruption — an approach with some similarities to ongoing research into safe AI interruptibility.

On the futarchy side, we are seeing interest ininterest rates as an objective function, a kind ofhybrid of futarchy and quadratic voting through voluntary coin locking as a governance algorithm, and various forms of moderated futarchy that give the futarchy enough power to prevent a majority collusion attack in a way that a democracy cannot, but otherwise leave the power to a voting process — all innovations that are at least worth the consideration of a group trying to use futarchy to build a world democracy DAO.

Another highly underappreciated solution is the use of governance algorithms that explicitly slow things down — the proposed DAO hard fork that may rescue the contained funds is only possible precisely because the DAO included a set of

rules that required every action to have a long delay time. Still another avenue that is starting to be explored is formal verification — using computer programs to automatically verify other computer programs, and make sure that they satisfy a set of claims about what the programs are supposed to do.

Formally proving “honesty” in the general case is impossible, due to thecomplexity of value problem, but we can make some partial guarantees to reduce risk. For example, we could formally prove that a certain kind of action cannot be taken in less than 7 days, or that a certain kind of action cannot be taken for 48 hours if the curators of a given DAO vote to flip a switch. In an AI context, such proofs could be used to prevent certain kinds of simple bugs in the reward function, that would result in a completely unintended behavior appearing to the AI to be of extremely high value. Of course, many other communities have been thinking about formal verification for many years already, but now it is being explored for a different use in a novel setting.

Meanwhile, one example of a concept promoted in the AI safety circles that may be highly useful to those building economic systems containing DAOs is superrational decision theories — essentially, ways to overcome prisoner’s dilemma situations by committing to run source code that treats agents which also commit to run that source code more favorably. One example of a move available to open-source agents that is not available to “black box” agents is the “values handshake” described in a short story by Scott Alexander: two agents can agree to both commit to maximize a goal which is the average of the two goals that they previously had. Previously, such concepts were largely science fiction, but now futarchy DAOs can actually do this

. More generally, a DAO may well be a highly effective means for a social institution to strongly commit to “running source code” that has particular properties.

“The DAO” is only the first in a series of many that will be launched over the course of this year and the next, and you can bet that all of the subsequent examples will learn heavily from the lessons of the first one, and each come up with different and innovative software code security policies, governance algorithms, curator systems, slow and phased bootstrap and rollout processes and formally verified guarantees in order to do its best to make sure that it can weather the cryptoeconomic storm.

Finally, I would argue that the biggest lesson to learn from the crypto community is that of decentralization itself: have different teams implement different pieces redundantly, so as to minimize the chance that an oversight from one system will pass through the other systems undetected. The crypto ecosystem is shaping up to be a live experiment comprising many challenges at the forefront of software development, computer science, game theory and philosophy, and the results, regardless of whether they make it into mainstream social applications in their present form or after several iterations that involve substantial changes to the core concepts, are welcome for anyone to learn from and see.