Behavioral Crypto-Economics: The Challenge n Incentive Design

In 2009, Satoshi Nakamoto designed Bitcoin to align the incentives of computers. blockchain projects, incentivizing humans using the same principles. We are assuming the efficient market hypothesis and the wisdom of the crowds. But what happens when the crowds are not that wise?

Elad Verbin · Follow
Published in Lunar Ventures
10 min read · Mar 16, 2018

Listen Share

Written by Elad Verbin and Al Esmail.

Bitcoin-style cryptoeconomic incentive design is a new economic design paradigm, that has already achieved incredible results, creating the first widespread digital currency. In recent years, this paradigm has become a “silver bullet”, used in far-reaching ways, promising to build amazing futuristic technology by employing incentives in planning and prediction (e.g. Gnosis, Augur); in social media (Steemit); in reputation, governance and self-organization (Colony, Boardroom, Democracy.Earth); in data collaboration and insights (Ocean, Numerai), and so on. However, in taking methods developed for simple systems and extending them into complex systems, we have taken a principle that was designed to incentivize algorithms to play by the rules, and applied it to people. We implicitly assumed people will make rational decisions, will take the most high-yield actions, and will collectively form an efficient market. However, behavioral economics argues that the wisdom of the crowd often isn’t all that wise, economically speaking. Behavioral economics teaches us that humans make a vast majority of their decisions based on simple mental shortcuts and rules of thumb, which are “usually correct” in common situations, but often turn out to be catastrophically wrong in edge cases (like when driving, voting, or transacting in complex financial markets); this is a well-documented phenomenon. In this post we argue that behavioral factors and psychology are not given enough consideration in cryptoeconomics. We think it imperative that experts in actual human economic behavior, such as public policy experts, behavioral economists and social scientists, be included in the teams designing cryptoeconomic systems, in order to ensure their long-term utility, viability and success.

1. Introduction: Bitcoin, Behavioral Economics, and Cryptoeconomics

Let’s start from the origins, Bitcoin. The concept of incentive design in blockchain originates from the original Bitcoin whitepaper by Satoshi Nakamoto, and is well summarized by Andreas Antonopoulos in his book and videos. Nakamoto used incentive design to achieve a previously-unattained goal: a scientifically-solid, secure, decentralized digital currency. Nakamoto's design incentivizes miners to secure the network and disincentivizes defection from the protocol's proper operations. Furthermore, it reasonably aligns the incentives of all stakeholders: miners, users, and developers contributing to the ecosystem. Its open-source nature ensures that an organized attack is not very lucrative, by enabling stakeholders to recognize the attack and defect to other chains. Seen another way, Nakamoto found a clever game-theoretic solution to the classic Byzantine Generals' Problem, by paying the generals a salary as long as they act honestly, but garnishing that salary if they are caught trying to cheat.[1]

On historical examination, Nakamoto's game theoretic assumptions are surprisingly mild. Bitcoin can be successfully disrupted only if 51% of the mining power cooperates and coordinates, aiming to disrupt it.[2] Crucially, Bitcoin's security does not depend on any "Homo Economicus" assumption that humans are ruthless optimizers and ultra-rational. Rather, even if people are lazy, and even if some malicious coalitions are formed, the system would still be secure. Compared to modern assumptions like those of Steemit or Augur, the assumptions of Bitcoin are much more realistic and uncontroversial. (Also, Bitcoin's security guarantees are comparatively quite strong, and have been mathematically proved; see footnote 2.)

Since 2009, incentive design has gotten much more sophisticated. Blockchain systems today find ever-cleverer ways to apply incentive structures to more complex systems:

- ZCash and other cryptocurrencies share the basic incentive structure pioneered by Bitcoin, as do Ethereum and other second-generation blockchain systems
- Gnosis, Augur and other prediction markets attempt to predict the future using a price discovery mechanism: incentivizing users to profit by trying to form accurate predictions of the future, and betting according to these predictions.
- Steemit incentivizes users to post interesting tidbits, and/or truthfully vote on the quality of other people's posts. Other reputation systems incentivize users to "upvote" reputable actors, thus creating a Blockchain analogue for humans' de-facto reputation systems. (None have proven themselves thus far.)
- Numerai incentivizes data scientists to devise good algorithms for trading in financial markets
- Futarchy incentivizes users to stake good decisions
- Ocean incentivizes users to stake good datasets and to provide added value to existing datasets (think Numerai meets Gnosis)
- Polkadot incentivizes stakeholders to make honest decisions in the network ("validators" and "collators"), to look for bad actors ("fishermen"), and to decide who is trustworthy ("nominators").

Incentive design is considered one of the killer features of blockchain systems

Overall, incentive design has spread to many exciting applications and is considered one of the killer features of blockchain systems. (More accurately, the killer feature is the ability to implement a fine-grained incentive system in a highly-scalable way, which supports tiny and large incentives alike). This is captured in the spirit of writing by the most prominent blockchain innovators. Trent McConaghy writes in a recent blog post:

> *'The blockchain community understands that blockchains can help* **align incentives** *among a tribe of token holders. Each token holder has* ***skin in the game****. But the benefit is actually more general than simply* aligning *incentives: you can* design *incentives of your choosing, by giving them block rewards. Put another way:* **you can get people to do stuff, by rewarding them with tokens.** *Blockchains are* ***incentive machines****.*
>
> *I see this as a* **superpower**. *The* block rewards function *defines what you want network participants to do. Then the question is: what do you want people in your network to do? It has a crucial corollary: how well can you* communicate that intent *to the machines? This is a devilish detail. Do we really know how to design incentives?'*[3]

This vision is powerful and inspiring. But it might raise more questions than it resolves. Below we explore McConaghy's question: "*Do we really know how to design incentives?*". We answer it with a resounding "No", and try to offer some ways to turn this "No" to a "Maybe".

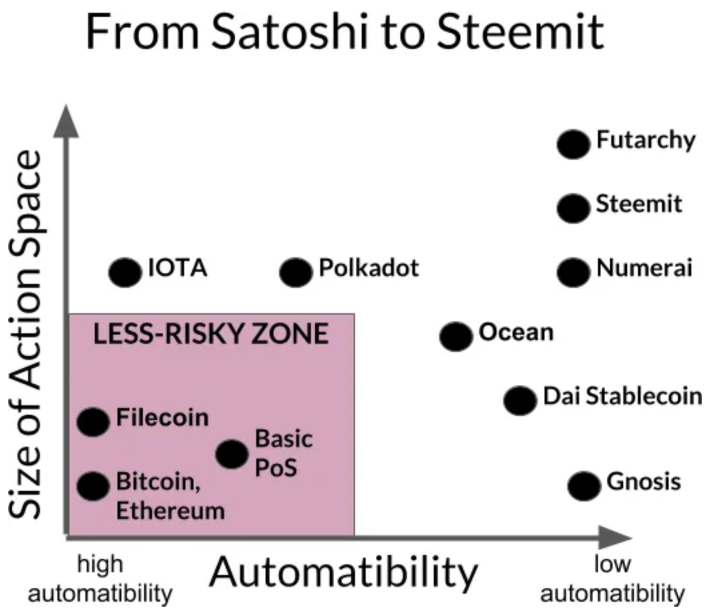

Figure 1. Qualitative scatter plot of decentralized systems by their automatibility and size of their Action Space

2. From Satoshi to Steemit

We will now analyze some ways in which incentive design has been extended since the original Bitcoin mold. In Figure 1 we qualitatively map the current landscape of highly-extended incentive design on a two-axis system.[4]

- The "automatibility' axis describes how much manual work is needed by the (human) stakeholders to follow the incentives. In a nutshell: In Bitcoin, computers make decisions, while in Steemit, humans make decisions. In detail: Bitcoin miners can follow incentives without lifting a finger, just by letting the computer mine honestly. Augur and Steemit, on the other hand, rely on humans to manually make predictions, write posts, and to get rewarded for it. In the middle, Polkadot and Proof-of-Stake systems require humans to stake entities that they trust to not act against the system.
- The "size of action space" axis describes how many possible actions must be explored in order to maximize reward. In Steemit the possible actions are as wide as the number of possible good posts, while in Bitcoin the effective action space consists of just one action: "mine and validate honestly". In Numerai, a human is asked to design a good algorithm, and gets rewarded according to its performance.

> Blockchain incentive design needs
> public policy experts as well as scientists.

We thus see that incentive design has been extended to incentivize intrinsically-human actions (e.g. writing blog posts) in vast action spaces.

But there's a problem: the belief in the power of incentives — which is so widespread in crypto-economics — is predicated on the stellar track record of Bitcoin. The Bitcoin paradigm, which justifiably gives credence to the bottom-left portion of the chart, is now being extended so that even systems occupying the top-right portion of the chart enjoy the same aura of reliability. Systems like Steemit, all the way in the top-right, are treated as run-of-the-mill blockchain-based cryptoeconomic systems, without distinguishing them from Bitcoin in any way, despite the fact that there has been no evaluation of whether their underlying cryptoeconomic models pass muster in the field over time. This points to an absence of expertise in the multidisciplinary skills needed for practical incentive design. These skills include mechanism design, cryptography and engineering, but *also* behavioral economics and the humanities: an understanding of the complex, nuanced, illogical ways in which humans actually behave. Blockchain incentive design needs the public policy experts as well as the scientists.

The difficulty is that humans are not rational actors. In reality, humans diverge from optimal behaviors very often. A classic example is the Ultimatum Game: person A is given $100 and asked to offer some amount of it to person B. Person B then has to "accept", or to "reject". If B "rejects", both players go aways with nothing. If B "accepts", B gets what was offered, and A gets the remaining part. Game theory tells us that the rational strategy for player B is to always accept, even if player A offers 0.01$. However, in reality, people typically reject offers of less than $30. Furthermore, people playing as A usually offer at least $20–30. So we see that in empirical settings both players adopt strategies that are wildly suboptimal (in the game-theoretic sense). We might well expect such effects to replicate in the intricate settings of cryptoeconomic systems such as Steemit. And what keeps irrationality from invalidating the design principles relied upon in cryptoeconomic systems that involve a substantial human component?

In particular, applying this “behaviorist” thinking to the axes in the above diagram, we claim that:

- The less automatible an incentive system is, the harder it is to design, and the more it is exposed to human irrationalities such as confirmation bias, sunk cost fallacies, and various types of Groupthink.
- The larger the action space, the harder it is to optimize and make the right choice. Computers can often search large spaces, but the cost of this might be prohibitive. With humans the situation is much worse. Humans hate having to choose, so the larger the search space, the higher the Cognitive Deliberation Cost, and the worse the results.

We thus see that in systems outside the “less-risky space” (the pink region in the chart), the actors themselves have a hard time figuring out which actions maximize their profits. And the challenge of the system designers is much harder: as it becomes harder for players to decide on actions, it becomes exponentially harder for the designer to design a stable, predictable, system. Applying mechanism design to produce systems that run well “in the wild” is a very difficult task even under the best conditions — where the action-spaces of individual players are simple and easy to analyze and optimize. When the actions are hard to optimize, as they would be outside the “less-risky space”, the mechanism designer hardly stands a chance.

3. Building Robust Incentive Systems?

So far, we have established that the design of complex cryptoeconomic systems is a difficult and not-yet-understood task. Thus cryptoeconomic systems will mostly start out broken, and have to undergo repeated iterations of improvements. But what’s wrong with that? Isn’t that the way technical systems have always worked? From the invention of fire and writing, through state-building and the space race, and onto computing, software and the internet — all of these start out deeply flawed, and become better over time.

Unfortunately, the situation with cryptocurrencies and blockchain systems is not as easily amenable to iterative improvements as other, historical, technologies. Firstly, blockchain systems are, by design, difficult to change once deployed. Any change (e.g. a fork) needs the blessing of most stakeholders, and thus any change has to fight entrenched interests (as exemplified by Bitcoin’s toxic Block Size Debate). Secondly, the weaknesses, vulnerabilities and design flaws that invariably exist in complex blockchain systems, are often not exposed until the systems are widely adopted. By then, multi-billion dollar market capitalizations and network effects make entrenched interests stronger and harder to fight. Whenever a community of users benefits from a design flaw, they will claim that it is “a feature, not a bug”, as incentives dictate. This is a well-known public policy conundrum, which will manifest in decentralized systems as they achieve prominence.

Adding fuel to the fire, the current exponential-growth regime may be hiding many of these faults. The behavior of a system whose tokens exponentially increase in value is much less adversarial and more amicable than when growth trickles to a stop. This means that as blockchain value growth slows down, we will see pull-down effects that may create a vicious cycle. Dysfunctional governance might make this cycle more vicious still..

In summary: good incentive systems are notoriously difficult to create under the best of circumstances. Under Blockchain systems — where code is law, etched in stone, and deployed to a wide community of pseudonymous stakeholders — good incentive systems are all the harder to get right. The success of Bitcoin should not make us complacent and optimistic: our incentive design paradigms needs to be scrupulous, slow growing, with a framework of checks and balances. If we design our new economies badly, then like in the last financial crisis, as the system begins to falter, it will trigger cascading effects that lengthen and deepen the fall. It is sobering to think that blockchain economies could meet the same fate as the very systems they were meant to replace: emboldened by exponential growth and short-term profits, we turn a blind eye to the long-term unsustainability of the model. Historically this led to external interventions in the market (though a taxpayer bailout may not be as feasible given blockchain’s stateless nature.)

The oracles and prophets of this brave new world would be well advised to heed caution when trying to build an economy from scratch. There are mounds of past examples and historical data to learn from among the rubble of fallen economic systems.