Tuur Demeester Follow
Economist & investor, Editor in Chief at Adamant Research
Jul 12 · 8 min read

Critique of Buterin's "A Proof of Stake Design Philosophy"

In this article, I take issue with several of the claims made by Vitalik Buterin in his Dec 2016 article "A Proof of Stake Design Philosophy". My hope is that it sparks debate about proof-of-stake's high level design and about the proposed future of the Ethereum protocol.

1. "Cost of attack should exceed cost of defense" is illogical

This is a core building block for the argument that proof-of-stake (PoS) is 'more efficient' than proof-of-work (PoW), so important to review carefully.

Vitalik starts this argument by claiming that cryptography allows users to defend their data in a much more effective way than a castle or island owner can self-defend in the physical world. While it is true that cryptography changes the game of wealth and information protection, often enabling a level playing field, this is comparing apples to oranges. Yes, it's true that a medieval knight cannot crack a bitcoin wallet, but neither can a computer hacker effectively defend a castle. Cryptography is used in the real world, where private keys worth millions can be stolen with a $5 wrench attack.

Moreover, 'cost of attack' and 'cost of defense' are not abstract and fixed, but rather relative and dynamic phenomena: they depend on the subjective value of the thing that one is attacking or defending, and on the conviction of the actors involved. Cost is a *relative* phenomenon, it only becomes meaningful once compared to forgone utility, to the opportunities the actor is willing to miss out on in order to pursue a particular goal. In the case of an attacker-defender scenario, cost is also dynamic: if I'm facing an attacker with high commitment and huge resources, my potential cost of defense will be very high, and vice versa.

When discussing proof-of-work, Buterin claims that it goes against the 'cypherpunk spirit' because in this system, the "cost of attack and cost of defense are at a 1:1 ratio". This statement is misleading, because he is really only talking about what a 51% attacker could do to the very last blocks in the blockchain.

Attacks on Bitcoin where one tries to reverse historical transactions which are more than a few days old are expensive in the extreme. Let's imagine that the person who paid 10,000 BTC for a pizza in May 2010 is now an evil villain ('Pizza Man') and he wants to reverse that regrettable transaction. To succeed, he would need to somehow infiltrate and control a full 100% of all Bitcoin mining rigs and mine for a period of over 200 days (or a smaller +51% percentage for much longer) in order to roll back the chain far enough with valid proof-of-work. After the multi-billion mining equipment acquisition costs, the cost of running the Bitcoin network for 200 days would be over $700 million (7.5 TWh at 10 cents/KWh). Now, the cost of defense against anything less than the Pizza Man attack is hard to compute, because it suffices for competing Bitcoin miners to simply follow their economic self-interest and mine Bitcoins for their own account—the protection of the network against a myriad of possible attacks is a side-effect.

Given that knowledge, subjective value, and resources are spread unevenly in society (just like in nature), there will always be a tug of war between attackers and defenders—no matter which security mechanism one uses. To speak of a cost/defense ratio of 1:1 is quite meaningless in my opinion.

To return to cryptocurrencies: one can try to design transaction clearing algorithms that are different from proof-of-work, but all you end up doing is obfuscating the work that attackers must do to exploit the system, and making it harder to define how much and which kind of work defenders need to do to keep the ledger honest and complete. Like Paul Sztorc has stated (also echoed by Adam Back): "all proposed PoW alternatives should be labeled 'obscured proof-of-work'".

2. No, humans are not "quite good at consensus"

Vitalik asserts that a 51% attacker who reverted the transaction ledger in his favor would have a very hard time convincing the community that his chain is legitimate. The crowd would unmask him and quickly reach consensus to restore justice. He continues: "these social considerations are what ultimately protect any blockchain in the long term", and cites the stone money on the island of Yap as an example.

First of all, I don't think the stone money from Yap is a good example of the effectiveness of social consensus. We have virtually no information about the amount of fraud committed or prevented under the stone money system. Further, it is well known that mores, customs, rituals, and social pressure play a much larger role in tribal communities like on the tiny island of Yap, so it's not fair to assume that one can successfully operate a similar system of monetary coordination in society at large. And finally, the Yap 'social consensus ledger' became victim of at least two successful attacks. The first was when in 1874 the Irish-American captain David O'Keefe managed to use large amounts of cheaply produced stones as currency to gain power and wealth. The second documented attack on the Yap financial system happened when German traders confiscated the yap stones and instituted harsh capital controls.

So let's focus on Buterin's assertion that social consensus is a protection against resource driven attacks. In my opinion that is plain wrong. An actor with the assets to conduct a such an operation can target his attack on very few individuals, and can make it expensive for the community to undo the theft and restore justice. Or the attacker can strategically target a huge amount of users, making sure to only inflict a small amount of financial damage per user—so that the cost per individual to rally against the attacker is higher than the loss incurred by the attack.

The fable of “belling the cat” is about a group of mice who debate plans to nullify the threat of a marauding cat. Putting a bell around the cat’s neck seems like an obviously good solution, until one mouse asks who will volunteer… The story illustrates how ‘social consensus’ can seem easy in theory, but is often hard in practice.

Even in the rare case where people largely agree that a certain event is disruptive and undesirable, they often entirely disagree on how it should be dealt with. Markets are good at letting people pursue their personal goals in a voluntary way, but that’s about it. If a subset of people (or an individual) doesn’t like something, they can always exit. In the universe of cryptocurrency, that means they can hard-fork and create their own new currency, or soft-fork to impose more stringent rules upon themselves.

All too often the word ‘consensus’ is used as rhetorical tool to silence dissent. For example, again in ‘A Proof of Stake Design Philosophy’, Buterin makes the claim that if a collusion of validators take over a proof-of-stake chain, “the community can simply coordinate a hard fork and delete the offending validators’ deposits”. Given that TheDAO bailout passed by supposed ‘community consensus’ even though less then 6% of Ether in circulation voted on the matter in a process of under 2 weeks, it seems risky to ‘offend’ the wrong people in the ETH community.

In sum, when faced with resource driven attacks, real response consensus is nigh impossible to achieve. Long or short term, political systems are not sufficiently reliable to prevent fraud and theft. In the pursuit of social scalability we can encourage individual liberty and responsibility by using the tools of cryptography, engineering, and economic self-interest as sources of robustness—but what we can *not* count on is the idealistic concept of social consensus.

3. Unsubstantiated claim that PoS is more resilient than PoW

Buterin states the following: “if desired, the cost of a single 51% attack on proof of stake can certainly be set to be as high as the cost of a *permanent* [sic] 51% attack on proof of work, and the sheer cost and ineffectiveness of an attack should ensure that it is almost never attempted in practice.”

In other words, he implies that from a security point of view, Proof-of-Stake is much more robust than Proof-of-Work.

In comparing PoW with PoS, consider the following:

- Cryptocurrency mining designs are solutions to the problem of trust in systems with imperfect knowledge and unknown adversaries. Proof-of-work has applications in early modern money and in nature, where the handicap principle evolutionarily evolved to let animals prove the “honesty” or reliability of their signal. To my knowledge, proof-of-stake has no equivalent applications in either human history or biology.

- A PoW 51% attacker can significantly slow down the network, but even a single attempt to revert historical transactions requires a huge and long-running expense. In other words, the production of ledger history is extremely expensive and its disruption arguably even more so.

- Contrary to a PoW-chain absent a +51% cartel, it's mathematically proven that it is impossible to determine the "true" transaction history in a PoS blockchain without an additional source of trust. If a source of trust is always needed, a potential pandora's box of attack and centralization scenarios is opened. This is a seed of truth behind the joke that Ethereum plans to use "proof of Vitalik".

- In a naive PoS environment, an attacker can easily create many alternative histories of the ledger, making it cheap to try different strategies. This is known as as the "nothing at stake problem". Ethereum plans to solve this by destroying the bonded security deposit of malicious validators. SolidX's Bob McElrath makes the point that the strategy of 'economic punishment' of attackers is moot if the punishment itself can be forked away. Another criticism of bonded PoS, as recently voiced by BitTorrent creator Bram Cohen, is the question how one prevents honest stakers from being tricked into interacting with the network in a way that triggers the punishment that is supposed to protect them. (Think of it as the crypto equivalent of large scale swatting.) An alternative attack scenario, suggested by Galois Capital's Kevin Zhou, is one where the attacker tricks enough honest people onto his network, so that it becomes these honest peoples interest to support the attacking chain as the true chain.

Conclusion

While it is commendable that Buterin works to build his cryptocurrency design proposals from first principles, I believe his write up contains several flaws. He is confused about cost-defense trade-offs and makes unsubstantiated claims about work- versus stake-based security. He fails to provide convincing logical or historical proof of the efficacy of social consensus. And he claims proof-of-stake is more resilient without providing proof or arguments, and without acknowledging the numerous objections that have been raised by people of substantial pedigree. Buterin's article does not convince me that proof-of-stake has a sound philosophical foundation, nor that it's a viable stand-alone mechanism for securing public blockchains.

I am grateful for the feedback of Kevin Zhou, Afsheen Bigdeli, Lawrence Nahum, Tommaso Pellizzari, and Christian Lundkvist. All errors remain my own.

Disclosure: I have a short position in ETH/BTC (short Ether, long Bitcoin).