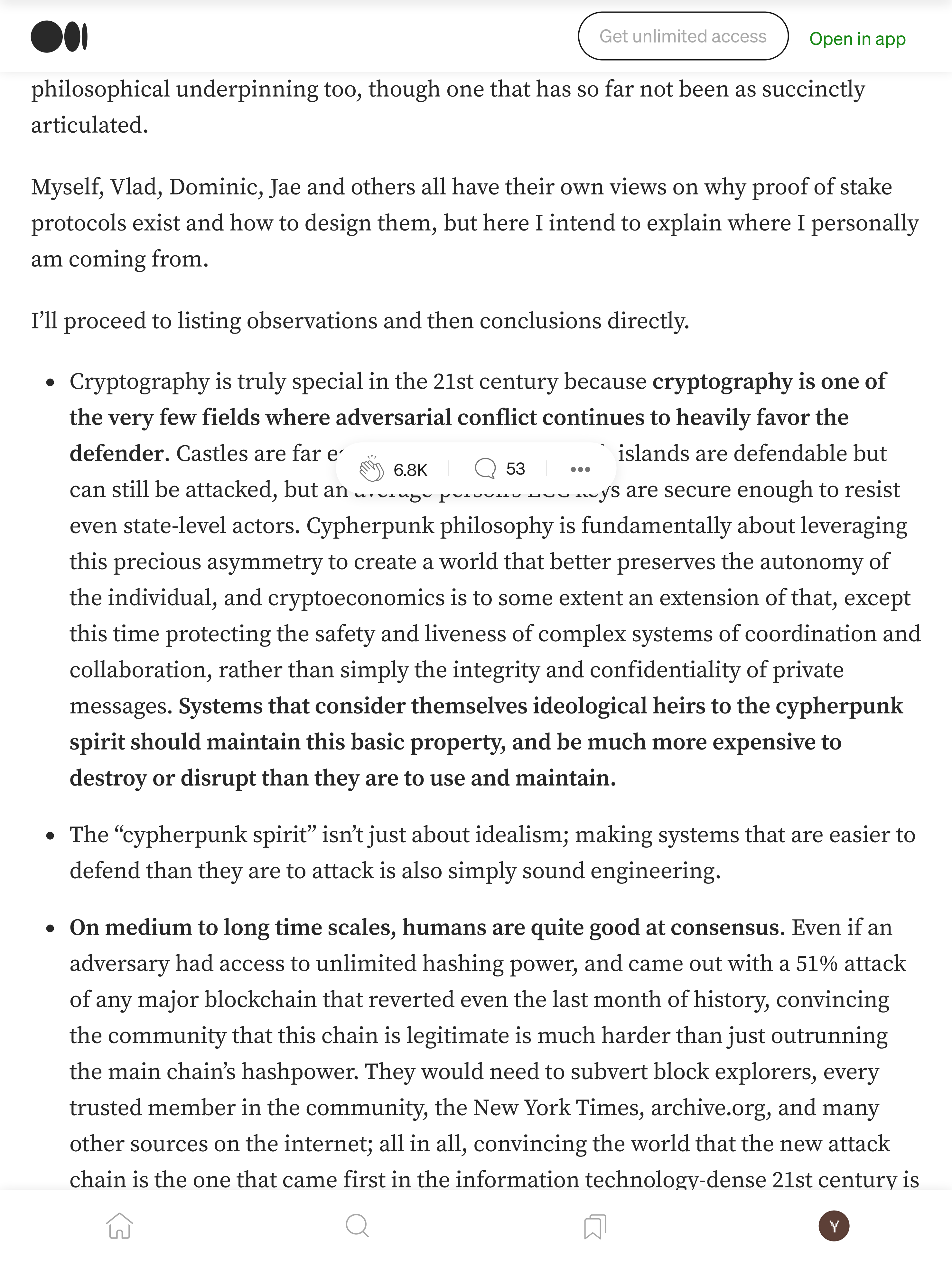
Vitalik Buterin Follow
Dec 31, 2016 · 7 min read · Listen

Save

A Proof of Stake Design Philosophy

Systems like Ethereum (and Bitcoin, and NXT, and Bitshares, etc) are a fundamentally new class of cryptoeconomic organisms — decentralized, jurisdictionless entities that exist entirely in cyberspace, maintained by a combination of cryptography, economics and social consensus. They are kind of like BitTorrent, but they are also not like BitTorrent, as BitTorrent has no concept of state — a distinction that turns out to be crucially important. They are sometimes described as decentralized autonomous corporations, but they are also not quite corporations — you can't hard fork Microsoft. They are kind of like open source software projects, but they are not quite that either — you can fork a blockchain, but not quite as easily as you can fork OpenOffice.

These cryptoeconomic networks come in many flavors — ASIC-based PoW, GPU-based PoW, naive PoS, delegated PoS, hopefully soon Casper PoS — and each of these flavors inevitably comes with its own underlying philosophy. One well-known example is the maximalist vision of proof of work, where "the" correct blockchain, singular, is defined as the chain that miners have burned the largest amount of economic capital to create. Originally a mere in-protocol fork choice rule, this mechanism has in many cases been elevated to a sacred tenet — see this Twitter discussion between myself and Chris DeRose for an example of someone seriously trying to defend the idea in a pure form, even in the face of hash-algorithm-changing protocol hard forks. Bitshares' delegated proof of stake presents another coherent philosophy, where everything once again flows from a single tenet, but one that can be described even more simply: shareholders vote.



philosophical underpinning too, though one that has so far not been as succinctly articulated.

Myself, Vlad, Dominic, Jae and others all have their own views on why proof of stake protocols exist and how to design them, but here I intend to explain where I personally am coming from.

I'll proceed to listing observations and then conclusions directly.

- Cryptography is truly special in the 21st century because **cryptography is one of the very few fields where adversarial conflict continues to heavily favor the defender.** Castles are far e islands are defendable but can still be attacked, but an average person's ECC keys are secure enough to resist even state-level actors. Cypherpunk philosophy is fundamentally about leveraging this precious asymmetry to create a world that better preserves the autonomy of the individual, and cryptoeconomics is to some extent an extension of that, except this time protecting the safety and liveness of complex systems of coordination and collaboration, rather than simply the integrity and confidentiality of private messages. **Systems that consider themselves ideological heirs to the cypherpunk spirit should maintain this basic property, and be much more expensive to destroy or disrupt than they are to use and maintain.**

- The "cypherpunk spirit" isn't just about idealism; making systems that are easier to defend than they are to attack is also simply sound engineering.

- **On medium to long time scales, humans are quite good at consensus.** Even if an adversary had access to unlimited hashing power, and came out with a 51% attack of any major blockchain that reverted even the last month of history, convincing the community that this chain is legitimate is much harder than just outrunning the main chain's hashpower. They would need to subvert block explorers, every trusted member in the community, the New York Times, archive.org, and many other sources on the internet; all in all, convincing the world that the new attack chain is the one that came first in the information technology-dense 21st century is

long term, regardless of whether or not the blockchain's community admits it (note that Bitcoin Core does admit this primacy of the social layer).

- However, a blockchain protected by social consensus alone would be far too inefficient and slow, and too easy for disagreements to continue without end (though despite all difficulties, it has happened); hence, **economic consensus serves an extremely important role in protecting liveness and safety properties in the short term.**

- Because proof of work security can only come from block rewards (in Dominic Williams' terms, it lacks two of the three Es), and incentives to miners can only come from the risk of them losing their future block rewards, **proof of work necessarily operates on a logic of massive power incentivized into existence by massive rewards.** Recovery from attacks in PoW is very hard: the first time it happens, you can hard fork to change the PoW and thereby render the attacker's ASICs useless, but the second time you no longer have that option, and so the attacker can attack again and again. Hence, the size of the mining network has to be so large that attacks are inconceivable. Attackers of size less than X are discouraged from appearing by having the network constantly spend X every single day. **I reject this logic because (i) it kills trees, and (ii) it fails to realize the cypherpunk spirit — cost of attack and cost of defense are at a 1:1 ratio, so there is no defender's advantage.**

- **Proof of stake breaks this symmetry by relying not on rewards for security, but rather penalties.** Validators put money ("deposits") at stake, are rewarded slightly to compensate them for locking up their capital and maintaining nodes and taking extra precaution to ensure their private key safety, but the bulk of the cost of reverting transactions comes from penalties that are hundreds or thousands of times larger than the rewards that they got in the meantime. **The "one-sentence philosophy" of proof of stake is thus not "security comes from burning energy", but rather "security comes from putting up economic value-at-loss".** A given block or state has $X security if you can prove that achieving an equal level of finalization

- Theoretically, a majority collusion of validators may take over a proof of stake chain, and start acting maliciously. However, (i) through clever protocol design, their ability to earn extra profits through such manipulation can be limited as much as possible, and more importantly (ii) if they try to prevent new validators from joining, or execute 51% attacks, then the community can simply coordinate a hard fork and delete the offending validators' deposits. **A successful attack may cost $50 million, but the process of cleaning up the consequences will not be *that* much more onerous than the geth/parity consensus failure of 2016.11.25**. Two days later, the blockchain and community are back on track, attackers are $50 million poorer, and the rest of the community is likely richer since the attack will have caused the value of the token to go *up* due to the ensuing supply crunch. *That's* attack/defense asymmetry for you.

- The above should not be taken to mean that unscheduled hard forks will become a regular occurrence; 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* 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.

- **Economics is not everything**. Individual actors may be motivated by extra-protocol motives, they may get hacked, they may get kidnapped, or they may simply get drunk and decide to wreck the blockchain one day and to hell with the cost. Furthermore, on the bright side, **individuals' moral forbearances and communication inefficiencies will often raise the cost of an attack to levels much higher than the nominal protocol-defined value-at-loss**. This is an advantage that we cannot rely on, but at the same time it is an advantage that we should not needlessly throw away.

- **Hence, the best protocols are protocols that work well under a variety of models and assumptions** — economic rationality with coordinated choice, economic rationality with individual choice, simple fault tolerance, Byzantine fault tolerance (ideally both the adaptive and non-adaptive adversary variants), Ariely/Kahneman-

from acting anti-socially, and anti-centralization incentives to discourage cartels from forming in the first place.

- **Consensus protocols that work as-fast-as-possible have risks and should be approached very carefully if at all**, because if the *possibility* to be very fast is tied to *incentives* to do so, the combination will reward very high and systemic-risk-inducing levels of **network-level centralization** (eg. all validators running from the same hosting provider). Consensus protocols that don't care too much how fast a validator sends a message, as long as they do so within some acceptably long time interval (eg. 4–8 seconds, as we empirically know that latency in ethereum is usually ~500ms-1s) do not have these concerns. A possible middle ground is creating protocols that can work very quickly, but where mechanics similar to Ethereum's uncle mechanism ensure that the marginal reward for a node increasing its degree of network connectivity beyond some easily attainable point is fairly low.

From here, there are of course many details and many ways to diverge on the details, but the above are the core principles that at least my version of Casper is based on. From here, we can certainly debate tradeoffs between competing values . Do we give ETH a 1% annual issuance rate and get an $50 million cost of forcing a remedial hard fork, or a zero annual issuance rate and get a $5 million cost of forcing a remedial hard fork? When do we increase a protocol's security under the economic model in exchange for decreasing its security under a fault tolerance model? Do we care more about having a predictable level of security or a predictable level of issuance? These are all questions for another post, and the various ways of *implementing* the different tradeoffs between these values are questions for yet more posts. But we'll get to it :)