Posted by Vitalik Buterin on August 21, 2014

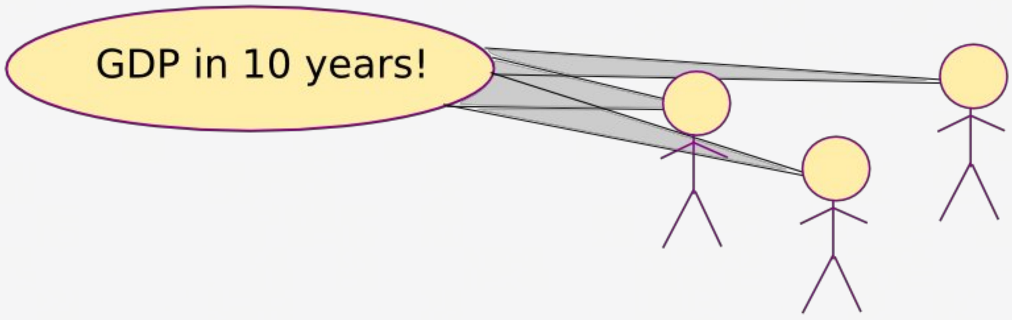
One of the more interesting long-term practical benefits of the technology and concept behind decentralized autonomous organizations is that DAOs allow us to very quickly prototype and experiment with an aspect of our social interactions that is so far arguably falling behind our rapid advancements in information and social technology elsewhere: organizational governance. Although our modern communications technology is drastically augmenting individuals' naturally limited ability to both interact and gather and process information, the governance processes we have today are still dependent on what may now be seen as centralized crutches and arbitrary distinctions such as "member", "employee", "customer" and "investor" - features that were arguably originally necessary because of the inherent difficulties of managing large numbers of people up to this point, but perhaps no longer. Now, it may be possible to create systems that are more fluid and generalized that take advantage of the full power law curve of people's ability and desire to contribute. There are a number of new governance models that try to take advantage of our new tools to improve transparency and efficiency, including liquid democracy and holacracy; the one that I will discuss and dissect today is futarchy.

The idea behind futarchy was originally proposed by economist Robin Hanson as a futuristic form of government, following the slogan: vote values, but bet beliefs. Under this system, individuals would vote not on whether or not to implement particular policies, but rather on a metric to determine how well their country (or charity or company) is doing, and then prediction markets would be used to pick the policies that best optimize the metric. Given a proposal to approve or reject, two prediction markets would be created each containing one asset, one market corresponding to acceptance of the measure and one to rejection. If the proposal is accepted, then all trades on the rejection market would be reverted, but on the acceptance market after some time everyone would be paid some amount per token based on the futarchy's chosen success metric, and vice versa if the proposal is rejected. The market is allowed to run for some time, and then at the end the policy with the higher average token price is chosen.

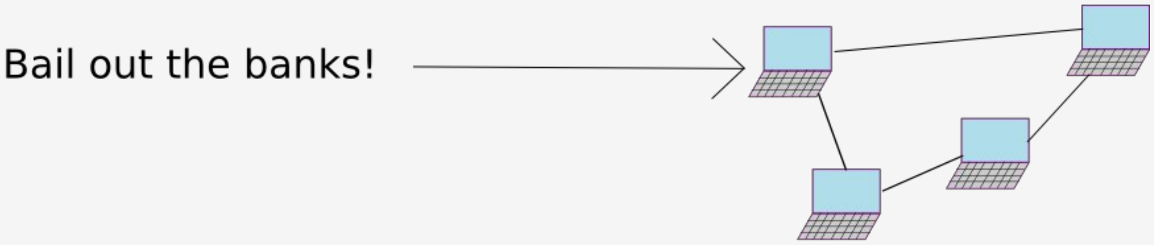
Our interest in futarchy, as explained above, is in a slightly different form and use case of futarchy, governing decentralized autonomous organizations and cryptographic protocols; however, I am presenting the use of futarchy in a national government first because it is a more familiar context. So to see how futarchy works, let's go through an example.

Suppose that the success metric chosen is GDP in trillions of dollars, with a time delay of ten years, and there exists a proposed policy: "bail out the banks". Two assets are released, each of which promises to pay $1 per token per trillion dollars of GDP after ten years. The markets might be allowed to run for two weeks, during which the "yes" token fetches an average price of $24.94 (meaning that the market thinks that the GDP after ten years will be $24.94 trillion) and the "no" token fetches an average price of $26.20. The banks are not bailed out. All trades on the "yes" market are reverted, and after ten years everyone holding the asset on the "no" market gets $26.20 apiece.

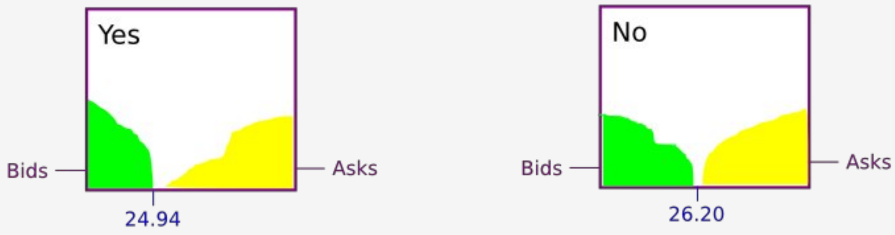
Step 0: choose a success metric and maturity duration



Step 1: create and publish proposal

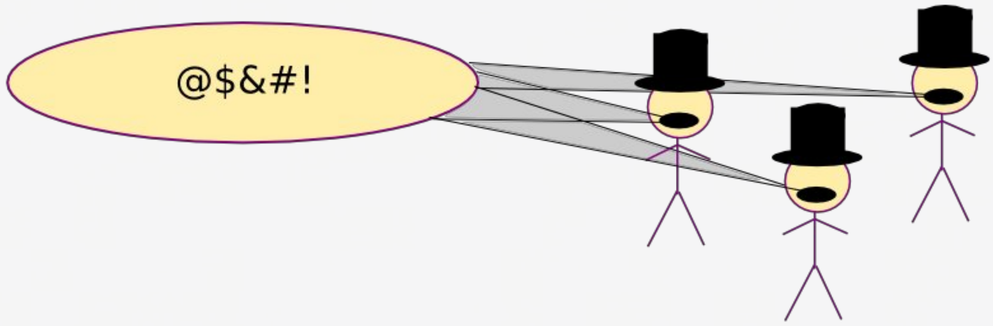


Step 2: set up prediction markets for "yes" and "no"

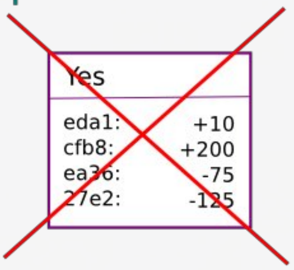


Note the average price of both over some period

Step 3: close both markets, implement the policy with the higher price

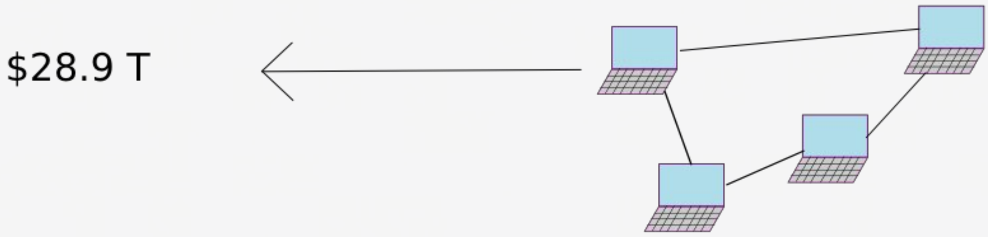


Step 4: revert all trades on losing market

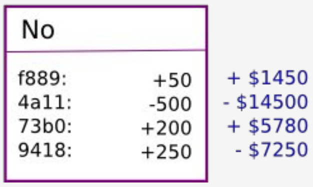


Yes	
eda1:	+10
cfb8:	+200
ea36:	-75
27e2:	-125

Step 5: wait for maturity, and measure success metric



Step 6: reward everyone on the winning market in proportion to how many tokens they have



No		
f889:	+50	+ $1450
4a11:	-500	- $14500
73b0:	+200	+ $5780
9418:	+250	- $7250

Typically, the assets in a futarchy are zero-supply assets, similar to Ripple IOUs or

BitAssets. This means that the only way the tokens can be created is through a derivatives market; individuals can place orders to buy or sell tokens, and if two orders match the tokens are transferred from the buyer to the seller in exchange for USD. It's possible to sell tokens even if you do not have them; the only requirement in that case is that the seller must put down some amount of collateral to cover the eventual negative reward. An important consequence of the zero-supply property is that because the positive and negative quantities, and therefore rewards cancel each other out, barring communication and consensus costs the market is actually free to operate.

Futarchy has become a controversial subject since the idea was originally proposed. The theoretical benefits are numerous. First of all, futarchy fixes the "voter apathy" and "rational irrationality" problem in democracy, where individuals do not have enough incentive to even learn about potentially harmful policies because the probability that their vote will have an effect is insignificant (estimated at 1 in 10 million for a US government national election); in futarchy, if you have or obtain information that others do not have, you can personally substantially profit from it, and if you are wrong you lose money. Essentially, you are literally putting your money where your mouth is.

Second, over time the market has an evolutionary pressure to get better; the individuals who are bad at predicting the outcome of policies will lose money, and so their influence on the market will decrease, whereas the individuals who are good at predicting the outcome of policies will see their money and influence on the market increase. Note that this is essentially the exact same mechanic through which economists argue that traditional capitalism works at optimizing the production of private goods, except in this case it also applies to common and public goods.

Third, one could argue that futarchy reduces potentially irrational social influences to the governance process. It is a well-known fact that, at least in the 20th century, the taller presidential candidate has been much more likely to win the election (interestingly, the opposite bias existed pre-1920; a possible hypothesis is that the switchover was caused by the contemporaneous rise of television), and there is the well-known story about voters picking George Bush because he was the president "they would rather have a beer with". In futarchy, the participatory governance process will perhaps encourage focusing more purely on proposals rather than personalities, and the primary activity is the most introverted and unsocial affair imaginable: poring over models, statistical analyses and trading charts.

90K
80K
70K
60K
50K
40K
30K
20K
10K
0K
1000
900
800
700
600
500
400
300
200
100
0
1 Dec
7
13
19
25
1 Jan
7
13
19
25
1 Feb
7
13
19
25

A market you would rather have a beer with

The system also elegantly combines public participation and professional analysis. Many people decry democracy as a descent to mediocrity and demagoguery, and prefer decisions to be made by skilled technocratic experts. Futarchy, if it works, allows individual experts and even entire analysis firms to make individual investigations and analyses, incorporate their findings into the decision by buying and selling on the market, and make a profit from the differential in information between themselves and the public - sort of like an information-theoretic hydroelectric dam or osmosis-based power plant. But unlike more rigidly organized and bureaucratic technocracies with a sharp distinction between member and non-member, futarchies allow anyone to participate, set up their own analysis firm, and if their analyses are successful eventually rise to the top - exactly the kind of generalization and fluidity we are looking for.

The opposition to futarchy is most well-summarized in two posts, one by Mencius Moldbug and the other by Paul Hewitt. Both posts are long, taking up thousands of words, but the general categories of opposition can be summarized as follows:

1. A single powerful entity or coalition wishing to see a particular result can continue buying "yes" tokens on the market and short-selling "no" tokens in order to push the token prices in its favor.
2. Markets in general are known to be volatile, and this happens to a large extent because markets are "self-referential" - ie. they consist largely of people buying because they see others buying, and so they are not good aggregators of actual information. This effect is particularly dangerous because it can be exploited by market manipulation.
3. The estimated effect of a single policy on a global metric is much smaller than the "noise" of uncertainty in what the value of the metric is going to be regardless of the policy being implemented, especially in the long term. This means that the prediction market's results may prove to be wildly uncorrellated to the actual delta that the individual policies will end up having.
4. Human values are complex, and it is hard to compress them into one numerical metric;

in fact, there may be just as many disagreements about what the metric should be as there are disagreements about policy now. Additionally, a malicious entity that in current democracy would try to lobby through a harmful policy might instead be able to cheat the futarchy by lobbying in an addition to the metric that is known to very highly correllate with the policy.

5. A prediction market is zero-sum; hence, because participation has guaranteed nonzero communication costs, it is irrational to participate. Thus, participation will end up quite low, so there will not be enough market depth to allow experts and analysis firms to sufficiently profit from the process of gathering information.

On the first argument, this video debate between Robin Hanson and Mencius Moldbug, with David Friedman (Milton's son) later chiming in, is perhaps the best resource. The argument made by Hanson and Friedman is that the presence of an organization doing such a thing successfully would lead to a market where the prices for the "yes" and "no" tokens do not actually reflect the market's best knowledge, presenting a massive profit-earning opportunity for people to put themselves on the opposite side of the attempted manipulation and thereby move the price back closer to the correct equilibrium. In order to give time for this to happen, the price used in determining which policy to take is taken as an average over some period of time, not at one instant. As long as the market power of people willing to earn a profit by counteracting manipulation exceeds the market power of the manipulator, the honest participants will win and extract a large quantity of funds from the manipulator in the process. Essentially, for Hanson and Friedman, sabotaging a futarchy requires a 51% attack.

The most common rebuttal to this argument, made more eloquently by Hewitt, is the "self-referential" property of markets mentioned above. If the price for "trillions of US GDP in ten years if we bail out the banks" starts off $24.94, and the price for "trillions of US GDP in ten years if we don't bail out the banks" starts off $26.20, but then one day the two cross over to $27.3 for yes and $25.1 for no, would people actually know that the values are off and start making trades to compensate, or would they simply take the new prices as an indicator of what the market thinks and accept or even reinforce them, as is often theorized to happen in speculative bubbles?

There is actually one reason to be optimistic here. Traditional markets may perhaps be often self-referential, and cryptocurrency markets especially so because they have no intrinsic value (ie. the only source of their value *is* their value), but the self-reference happens in part for a different reason than simply investors following each other like lemmings. The mechanism is as follows. Suppose that a company is interested in raising funds through share issuance, and currently has a million shares valued at $400, so a market cap of $400 million; it is willing to dilute its holders with a 10% expansion. Thus, it can raise $40 million. The market cap of the company is supposed to target the total amount of dividends that the company will ever pay out, with future dividends appropriately discounted by some interest rate; hence, if the price is stable, it means that the market expects the company to eventually release the equivalent of $400 million in total dividends in present value.

Now, suppose the company's share price doubles for some reason. The company can now raise $80 million, allowing it to do twice as much. Usually, capital expenditure has diminishing returns, but not always; it may happen that with the extra $40 million capital the company will be able to earn twice as much profit, so the new share price will be perfectly justified - even though the cause of the jump from $400 to $800 may have been manipulation or random noise. Bitcoin has this effect in an especially pronounced way; when the price goes up, all Bitcoin users get richer, allowing them to build more businesses, justifying the higher price level. The lack of intrinsic value for Bitcoin means that *the self-referential effect is the only effect* having influence on the price.

Prediction markets do not have this property at all. Aside from the prediction market itself, there is no plausible mechanism by which the price of the "yes" token on a prediction market will have any impact on the GDP of the US in ten years. Hence, the only effect by which self-reference can happen is the "everyone follows everyone else's judgement" effect. However, the extent of this effect is debatable; perhaps because of the very recognition that the effect exists, there is now an established culture of smart contrarianism in investment, and politics is certainly an area where people are willing to keep to unorthodox views. Additionally, in a futarchy, the relevant thing is not how high individual prices are, but which one of the two is *higher*; if you are certain that bailouts are bad, but you see the yes-bailout price is now $2.2 higher for some reason, you know that something is wrong so, in theory, you might be able to pretty reliably profit from that.

This is where we get to the crux of the real problem: it's not clear how you can. Consider a more extreme case than the yes/no bailouts decision: a company using a futarchy to determine how much to pay their CEO. There have been studies suggesting that ultra-high-salary CEOs actually do not improve company performance - in fact, much the opposite. In order to fix this problem, why not use the power of futarchy and the market decide how much value the CEO really provides? Have a prediction market for the company's performance if the CEO stays on, and if the CEO jumps off, and take the CEO's salary as a standard percentage of the difference. We can do the same even for lower-ranking executives and if futarchy ends up being magically perfect even the lowliest employee.

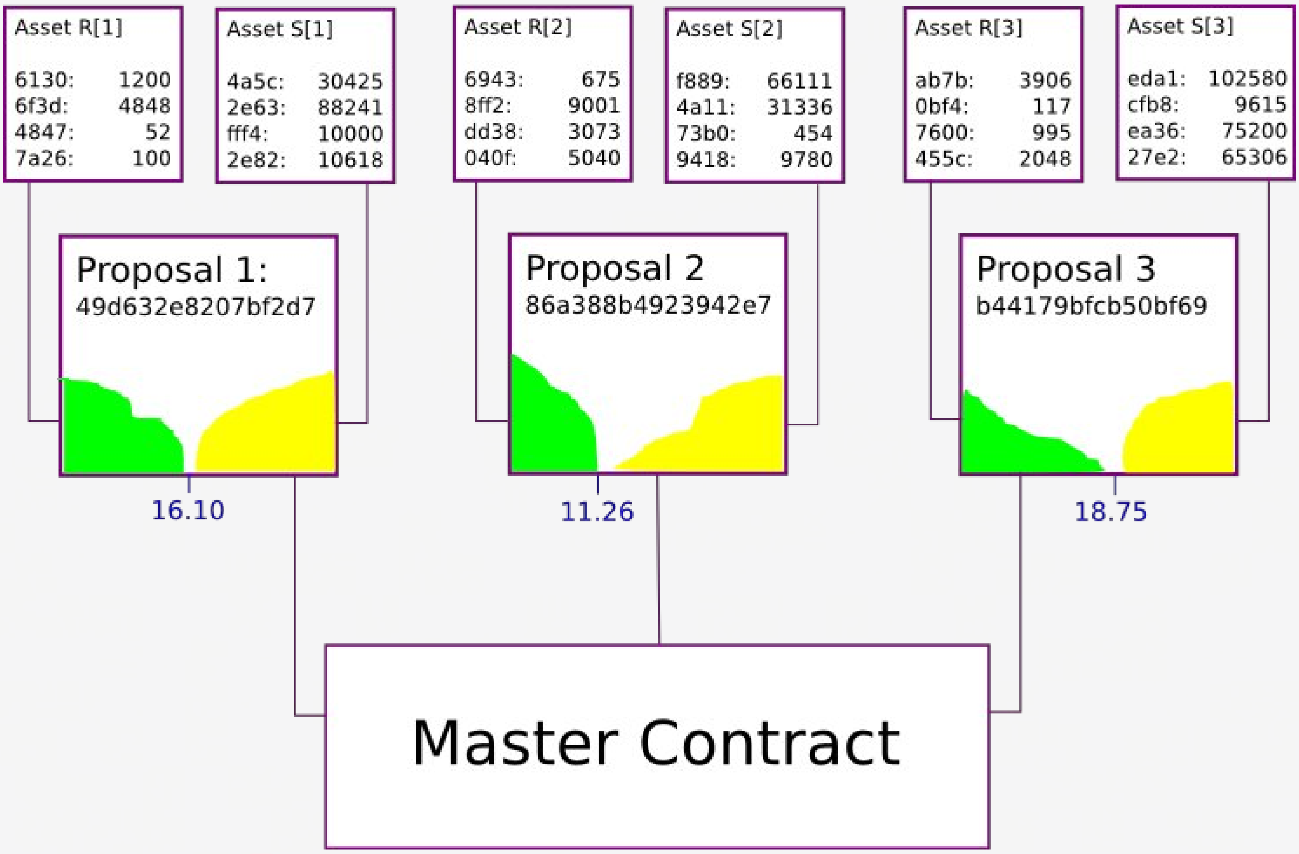
Now, suppose that you, as an analyst, predict that a company using such a scheme will have a share price of $7.20 in twelve months if the CEO stays on, with a 95% confidence interval of $2.50 (ie. you're 95% sure the price will be between $4.70 and $9.70). You also predict that the CEO's benefit to the share price is $0.08; the 95% confidence interval that you have here is from $0.03 to $0.13. This is pretty realistic; generally errors in measuring a variable are proportional to the value of that variable, so the range on the CEO will be much lower. Now suppose that the prediction market has the token price of $7.70 if the CEO stays on and $7.40 if they leave; in short, the market thinks the CEO is a rockstar, but you disagree. But how do you benefit from this?

The initial instinct is to buy "no" shares and short-sell "yes" shares. But how many of each? You might think "the same number of each, to balance things out", but the problem is that the chance the CEO will remain on the job is much higher than 50%. Hence, the "no" trades will probably all be reverted and the "yes" trades will not, so alongside shorting the CEO what you are also doing is taking a much larger risk shorting the company. If you knew the percentage change, then you could balance out the short and long purchases such that on net your exposure to unrelated volatility is zero; however, because you don't, the risk-to-reward ratio is very high (and even if you did, you would still be exposed to the variance of the company's global volatility; you just would not be biased in any particular direction).

From this, what we can surmise is that futarchy is likely to work well for large-scale decisions, but much less well for finer-grained tasks. Hence, a hybrid system may work better, where a futarchy decides on a political party every few months and that political party makes decisions. This sounds like giving total control to one party, but it's not; note that if the market is afraid of one-party control then parties could voluntarily structure themselves to be composed of multiple groups with competing ideologies and the market would prefer such combinations; in fact, we could have a system where politicians sign up as individuals and anyone from the public can submit a combination of politicians to elect into parliament and the market would make a decision over all combinations (although this would have the weakness that it is once again more personality-driven).

All of the above was discussing futarchy primarily as a political system for managing government, and to a lesser extent corporations and nonprofits. In government, if we apply futarchy to individual laws, especially ones with relatively small effect like "reduce the duration of patents from 20 years to 18 years", we run into many of the issues that we described above. Additionally, the fourth argument against futarchy mentioned above, the complexity of values, is a particular sore point, since as described above a substantial portion of political disagreement is precisely in terms of the question of what the correct values are. Between these concerns, and political slowness in general, it seems unlikely that futarchy will be implemented on a national scale any time soon. Indeed, it has not even really been tried for corporations. Now, however, there is an entirely new class of entities for which futarchy might be much better suited, and where it may finally shine: DAOs.

To see how futarchy for DAOs might work, let us simply describe how a possible protocol would run on top of Ethereum:



1. Every round, `T` new DAO-tokens are issued. At the start of a round, anyone has the ability to make a proposal for how those coins should be distributed. We can simplify and say that a "proposal" simply consists of "send money to this address"; the actual plan for how that money would be spent would be communicated on some higher-level channel like a forum, and trust-free proposals could be made by sending to a contract. Suppose that `n` such proposals, `P[1]` ... `P[n]`, are made.
2. The DAO generates `n` pairs of assets, `R[i]` and `S[i]`, and randomly distributes the `T` units of each type of token in some fashion (eg. to miners, to DAO token holders, according to a formula itself determined through prior futarchy, etc). The DAO also provides `n` markets, where market `M[i]` allows trade between `R[i]` and `S[i]`.
3. The DAO watches the average price of `S[i]` denominated in `R[i]` for all markets, and lets the markets run for `b` blocks (eg. 2 weeks). At the end of the period, if market `M[k]` has the highest average price, then policy `P[k]` is chosen, and the next period begins.
4. At that point, tokens `R[j]` and `S[j]` for `j != k` become worthless. Token `R[k]` is worth `m` units of some external reference asset (eg. ETH for a futarchy on top of

Ethereum), and token `S[k]` is worth `z` DAO tokens, where a good value for `z` might be 0.1 and `m` self-adjusts to keep expenditures reasonable. Note that for this to work the DAO would need to also sell its own tokens for the external reference asset, requiring another allocation; perhaps `m` should be targeted so the token expenditure to purchase the required ether is `zT`.

Essentially, what this protocol is doing is implementing a futarchy which is trying to optimize for the token's price. Now, let's look at some of the differences between this kind of futarchy and futarchy-for-government.

First, the futarchy here is making only a very limited kind of decision: to whom to assign the `T` tokens that are generated in each round. This alone makes the futarchy here much "safer". A futarchy-as-government, especially if unrestrained, has the potential to run into serious unexpected issues when combined with the fragility-of-value problem: suppose that we agree that GDP per capita, perhaps even with some offsets for health and environment, is the best value function to have. In that case, a policy that kills off the 99.9% of the population that are not super-rich would win. If we pick plain GDP, then a policy might win that extremely heavily subsidizes individuals and businesses from outside relocating themselves to be inside the country, perhaps using a 99% one-time capital tax to pay for a subsidy. Of course, in reality, futarchies would patch the value function and make a new bill to reverse the original bill before implementing any such obvious egregious cases, but if such reversions become too commonplace then the futarchy essentially degrades into being a traditional democracy. Here, the worst that could happen is for all the N tokens in a particular round to go to someone who will squander them.

Second, note the different mechanism for how the markets work. In traditional futarchy, we have a zero-total-supply asset that is traded into existence on a derivatives market, and trades on the losing market are reverted. Here, we issue positive-supply assets, and the way that trades are reverted is that the entire issuance process is essentially reverted; both assets on all losing markets become worth zero.

The biggest difference here is the question of whether or not people will participate. Let us go back to the earlier criticism of futarchy, that it is irrational to participate because it is a zero-sum game. This is somewhat of a paradox. If you have some inside information, then you might think that it is rational to participate, because you know something that other people don't and thus your expectation of the eventual settlement price of the assets is different from the market's; hence, you should be able to profit from the difference. On the other hand, if everyone thinks this way, then even some people with inside information will lose out; hence, the correct criterion for participating is something like "you should participate if you think you have better inside information than everyone else participating". But if everyone thinks this way then the equilibrium will be that no one participates.

Here, things work differently. People participate by default, and it's harder to say what not participating is. You could cash out your `R[i]` and `S[i]` coins in exchange for DAO tokens, but then if there's a desire to do that then `R[i]` and `S[i]` would be undervalued and there would be an incentive to buy both of them. Holding only `R[i]` is also not non-participating; it's actually an expression of being bearish on the merits of policy `P[i]`; same with holding only `S[i]`. In fact, the closest thing to a "default" strategy is holding whatever `R[i]` and `S[i]` you get; we can model this prediction market as a zero-supply market plus this extra initial allocation, so in that sense the "just hold" approach *is* a default. However, we can argue that the barrier to participation is much lower, so participation will increase.

Also note that the optimization objective is simpler; the futarchy is not trying to mediate the rules of an entire government, it is simply trying to maximize the value of its own token by allocating a spending budget. Figuring out more interesting optimization objectives, perhaps ones that penalize common harmful acts done by existing corporate entities, is an unsolved challenge but a very important one; at that point, the measurement and metric manipulation issues might once again become more important. Finally, the actual day-to-day governance of the futarchy actually does follow a hybrid model; the disbursements are made once per epoch, but the management of the funds within that time can be left to individuals, centralized organizations, blockchain-based organizations or potentially other DAOs. Thus, we can expect the differences in expected token value between the proposals to be large, so the futarchy actually will be fairly effective - or at least more effective than the current preferred approach of "five developers decide".

So what are the practical benefits of adopting such a scheme? What is wrong with simply having blockchain-based organizations that follow more traditional models of governance, or even more democratic ones? Since most readers of this blog are already cryptocurrency advocates, we can simply say that the reason why this is the case is the same reason why we are interested in using cryptographic protocols instead of centrally managed systems - cryptographic protocols have a much lower need for trusting central authorities (if you are not inclined to distrust central authorities, the argument can be more accurately rephrased as "cryptographic protocols can more easily generalize to gain the efficiency, equity and informational benefits of being more participatory and inclusive without leading to the consequence that you end up trusting unknown individuals"). As far as social consequences go, this simple version of futarchy is far from utopia, as it is still fairly similar to a profit-maximizing corporation; however, the two important improvements that it does make are (1) making it harder for *executives* managing the funds to cheat both the organization and society for their short-term interest, and (2) making governance radically open and transparent.

However, up until now, one of the major sore points for a cryptographic protocol is how the protocol can fund and govern itself; the primary solution, a centralized organization with a one-time token issuance and presale, is basically a hack that generates *initial* funding and *initial* governance at the cost of *initial* centralization. Token sales, including our own Ethereum ether sale, have been a controversial topic, to a large extent because they introduce this blemish of centralization into what is otherwise a pure and decentralized cryptosystem; however, if a new protocol starts off issuing itself as a futarchy from day one, then that protocol can achieve incentivization without centralization - one of the key breakthroughs in economics that make the cryptocurrency space in general worth watching.

Some may argue that inflationary token systems are undesirable and that dilution is bad; however, an important point is that, if futarchy works, this scheme is guaranteed to be at least as effective as a fixed-supply currency, and in the presence of a nonzero quantity of potentially satisfiable public goods it will be strictly superior. The argument is simple: it is always possible to come up with a proposal that sends the funds to an unspendable address, so any proposal that wins would have to win against that baseline as well.

So what are the first protocols that we will see using futarchy? Theoretically, any of the higher-level protocols that have their own coin (eg. SWARM, StorJ, Maidsafe), but without their own blockchain, could benefit from futarchy on top of Ethereum. All that they would need to do is implement the futarchy in code (something which I have started to do already), add a pretty user interface for the markets, and set it going. Although technically every single futarchy that starts off will be exactly the same, futarchy is Schelling-point-dependent; if you create a website around one particular futarchy, label it "decentralized insurance", and gather a community around that idea, then it will be more likely that that particular futarchy succeeds if it actually follows through on the promise of decentralized insurance, and so the market will favor proposals that actually have something to do with that particular line of development.

If you are building a protocol that will have a blockchain but does not yet, then you can use futarchy to manage a "protoshare" that will eventually be converted over; and if you are building a protocol with a blockchain from the start you can always include futarchy right into the core blockchain code itself; the only change will be that you will need to find something to replace the use of a "reference asset" (eg. 2^{64} hashes may work as a trust-free economic unit of account). Of course, even in this form futarchy cannot be guaranteed to work; it is only an experiment, and may well prove inferior to other mechanisms like liquid democracy - or hybrid solutions may be best. But experiments are what cryptocurrency is all about.

← →