Wanted to share some thoughts and get feedback on how Maker can develop GAIT more effectively while supporting the greater effort to solve the problem of AI alignment (originally posted here):

With the development of Maker's Governance AI Tools (GAIT) we have an opportunity to support the wider ecosystem through open source AI. Perhaps a more significant contribution to the public good however is the potential to showcase how crypto can help in solving AI alignment. It is also an opportunity to experiment with innovative models for subDAO funding and governance.

Broadly speaking, the problem of AI alignment is the challenge to ensure that, as AI tech advances, the superintelligent system works in a manner that is consistent with human values and interests. An AI system that is significantly more capable than humans and acts against human interests can have catastrophic results.

So how can Maker contribute on this front? Let's start with a thought experiment: imagine you have a government that is supposed to represent the people and their interests. Yet, funding for the government comes from a particular industry (at this point it doesn't matter which industry — oil, pharma, defense, big tech, and so on). Can such a government faithfully represent the interests of the people, or is it likely to be in a conflict of interests?

Even if politicians in such a government have the most noble intentions, the problem is their incentives. Every time the government needs to choose between policy that is in the interest of the people and policy that benefits its source of funding (thus benefiting itself financially) they have a problem. Either the government won't be able to fund itself sustainably or it won't faithfully represent the people's interests.

So what does this have to do with GAIT and AI alignment? Similar to the government thought experiment, both AI developers and the AI system itself have a potential misalignment between their source of funding and the interests of the community.

The fundamental question is, if the funding for developing AI doesn't come directly from the community that the AI is meant to serve, why would the economic incentives of developers align with the community? And if developers' incentives are misaligned, why would their product - GAIT - be aligned?

The question is not merely whether the source of funding has the community's best interest in mind. Even if it does, it still cannot possibly know all the community's priorities. So as long as developers don't report to the community, but rather to an extraneous party, their work would not be aligned with the community's priorities. What's more, if the community doesn't fund the work directly it would have less of a stake in the success of the project.

It is only fair that those who stand to benefit the most from GAIT would also have a bigger stake in the development process. We would therefore want the funding to come from all members of the community. We would also want each member's contribution to be proportionate to their stake in the community.

This tells us that it's not enough for the funding to come from a common DAO treasury. Such a model would certainly be better than an external funding source, but it would not create the proper incentives within the community to effectively oversee AI development.

The funding model that can effectively align developer–community incentives is token inflation. The concept here is as follows: since GAIT is expected to improve the internal dynamics of the system, it is likely to attract more participants and more investment to the ecosystem. The increase in demand will therefore lead to growth in the value of the subDAO's native token. This means that as long as token inflation (supply growth) is smaller than the increase in demand induced by improvements due to GAIT, the community would financially benefit from investing in AI development.

In this model, every community member would likely be more engaged in the GAIT development process, instead of passively observing development from the sidelines. They would want to choose the most competent developers, as well as take an active role in supporting AI integration. Meanwhile, developers would have to report to the community, which means that their efforts would be aligned with developing a system that prioritizes the community's needs.

The actual payment for the work can be structured in a way that incentivizes more effective development; releasing payments based on milestones, AI capabilities, and so on. This would ensure that the subDAO can keep track of the development process and maintain the value of its token.

The idea here is that the demand for the token partly depends on expectations. Why does this matter? Until the AI system is fully operational the community will not see immediate benefits from the system. And yet, developers need to be paid for their work. If payments for development increase the token supply it would lead to token devaluation. However, if everyone can see that payments are only made once milestones are met, the expectation for future benefits from AI would counteract any sell pressure. This means that, despite a lack of immediate benefits, token inflation in such a model should not lower the value of the subDAO's token. In fact, it may actually increase the value of the token. That is because the community (and outside observers) will be able to see that the funding process is running smoothly and progress is being made on AI development.

So now we have a mechanism where AI development can be aligned with the values and interests of the community. Which means that the AI system would be designed to benefit the community the most and not some extraneous interests. But that doesn't necessarily mean that we've achieved AI alignment yet.

The AI system at hand is based on reward functions, where the AI gets a score or feedback based on its ongoing performance. This means that if we want AI to stay aligned with the values and interests of the community, the “reward” for its performance, as well as ongoing operating costs, should also come from token inflation.

Such a design would ensure that the community stays engaged in fairly evaluating the performance of the system and is interested in its success. This design would also require currency sinks to maintain the value of the token.

There are still many unanswered questions in the proposed design; how will the subDAO make decisions? How will members come to consensus on policy? How do we prevent bad actors from manipulating this system?

These questions certainly need to be resolved. And yet, we can already see how such a design can have far-reaching implications beyond Maker. It can serve as a prototype to developing user and community-centric open source AI models that are self-sustainable and don’t require external funding. Models that, when built at scale, may be able to effectively compete with proprietary AI (in terms of capacity and development budget). It can also showcase how crypto (and Maker) can have use cases that serve the public interest and tackle some of the most important challenges we have today.