SATYA WHITEPAPER

Ilya Grigorenko and Theo Grigorenko

Draft

April 2024

ABSTRACT

In today's interconnected digital world, we witness how readily accessible information, free or nearly free, can become tainted with falsehoods and "fake news," to the detriment of humanity. Bogus claims and falsified facts are much easier to create than to verify. They rapidly spread among uninformed or misinformed individuals, perpetuating themselves and negatively affecting people and their decisions. There is currently an extraordinary demand for sources of credible, independently verified information. Such verified information holds relatively high internal value and can be consumed by humans as well as used to train various AI algorithms. This project aims to address the problem of "fake news" by providing a platform and a theoretically provable framework to generate minimally biased information ratings and ratings of the information providers.

The proposed solution utilizes a distributed, blockchain-based platform that supports independence, protects from censorship, and minimizes malicious external interference. The platform ensures that verified, non-anonymous participants from diverse backgrounds and with various levels of expertise can participate in confirming the veracity of news, statements, and facts. The verification process relies on a consensus among optimally weighted opinions of experts. Active and successful participation results in increased weights and monetary rewards for the participants, while demonstrated poor performance or malicious behavior is penalized. The framework provides a robust defense against various types of reputation attacks on experts, using a smart weighting algorithm and incentives based on market-like mechanisms.

INTRODUCTION TO THE PROBLEM

Over the last 30 years, disseminating various types of information through consumer networks has become a highly profitable business. However, not all information is equally trustworthy, especially when it relates to recent events. Studies have shown that false information spreads much faster than accurate information, creating financial incentives to produce fake stories[1]. Centralized social media networks and search engines have exacerbated the spread of fake news by rewarding users for having their beliefs validated. Unfortunately, effective mechanisms to halt the replication of false claims are lacking, except for direct intervention by major social network corporations or legal action against the originators of fake news, which is challenging in the age of anonymous internet access and easily falsifiable social network accounts[2]. Most importantly, news corporations, social network platforms, and their owners have their own financial interests, political preferences, and affiliations; therefore, they cannot be considered fully unbiased platforms or reliable information curators. One example of biased censorship was when Facebook suppressed the Hunter Biden laptop story ahead of the 2020 election. Mark Zuckerberg admitted the challenge of content moderation in a rare extended media interview on the Joe Rogan podcast.

Advancements in machine learning and artificial intelligence have made it easier to create misleading texts, pictures, audio, and video materials of unprecedented quality. With the current trends, one can anticipate that in the near future, the amount of AI-generated content will surpass that created by humans, further contaminating the internet as a source of information. Thus, it is paramount to create a self-regulated platform that penalizes insufficient verification of published news or intentional misinformation while incentivizing members who consistently provide high-quality information.

[1] Study: On Twitter, false news travels faster than true stories: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[2] Fake Polls, Real Consequences: The Rise of Fake Polls and the Case for Criminal Liability (missouri.edu). https://scholarship.law.missouri.edu/cgi/viewcontent.cgi?article=4418&context=mlr

PROPOSED SOLUTION

Similar to equity markets where the market generally rewards external investors' ability to identify reliable and financially successful companies—or unreliable ones in the case of short selling—a startup company is often evaluated based on the perceived value of the idea, the past record of the founder, their established expertise, and how deeply the founder is involved and invested in the startup. This holds particularly true in the long run.

Our proposed solution employs a similar approach, requiring fact/news makers as well as voting experts to back their claimed correct information with their own money, literally putting their money where their mouth is. We propose a new platform for a decentralized information and news market that employs blockchain technology to ensure the immutability of confirmed facts. The amount of money a newsmaker is willing to place behind their own story serves as an immediate indicator of their trustworthiness, thereby supporting their honesty through costly signaling. Any inherent bias, negligence in fact-checking, or the dissemination of outright falsehoods by the newsmaker is likely to be penalized by other platform members who can contest the story's credibility. All positions are formalized through standardized contracts, featuring upvote/downvote options that fall within specific percentage ranges at predetermined times post-publication. To prevent manipulation and various 'pump-and-dump' schemes, each authorized participant is limited to one vote and is prohibited from voting against their own published facts.

After the contract expires, all sides are settled among the participants, and the proceeds are distributed in the platform's internal digital currency. The participants' profits are calculated based on the timing of their entry: each participant can gain a proportional share of the money from opposing parties placed after their own, plus one position placed immediately before their own. This time-based ordering incentivizes early entry and discourages unfair last-minute positions that would almost guarantee a win at the expense of earlier participants. We have considered standardized time frames for positions following the publication of a news story, such as 1 hour, 1 day, 1 week, 1 month, 1 year, and 10 years.

For those interested in more precise outcomes, we offer the option to predict a specific range for the aggregated support for the fact at the time of the contract's expiration. In the simplest case, one could use five bins that correspond to different levels of support:

- High confidence false: 0-20%
- Likely false: 20%-40%
- Uncertain: 40%-60%
- Likely true: 60%-80%
- High confidence true: 80%-100%

Alternatively, one can also consider unevenly distributed bins:

- High confidence false: 0-5%
- Likely false: 5%-35%
- Uncertain: 35%-75%
- Likely true: 75%-95%
- High confidence true: 95%-100%

However, this type of contracts would require a more sophisticated proceeds-sharing scheme.

Registration and ID Verification

We have created a version of the app for Android OS smartphones. We will create and verify an iPhone version as well, along with a regular website-based gateway. Registered participants with verified IDs may place comments, take sides in the fact verification process, and publish their own material.

Name verification goes through a self-sovereign identity process (still to be completed). The identification process will be underpinned by established blockchain solutions used in the verification of digital identities, such as the Verifiable Credentials Data Model 1.0 (as per w3.org), Sovrin, and possibly others.

Publication Categories: Facts, Opinions and Beliefs

The current prototype of the platform categorizes publications into three main groups: beliefs, personal opinions, and facts. Only published information under the "fact" category is subject to monetary renumeration/penalties for fact verification, and any publication labeled as a fact will include an immutable link to the originator of the information. Publications containing verified information are marked as such with a green mark.

In the future, we plan to employ an AI-based monitoring tool that will suggest the appropriate classifications for each publication. Incorrect categorizations, such as mislabeling opinions or beliefs as facts, can also be contested through a voting mechanism.

If the vote concludes the categorization is incorrect, it may result in a decrease in the author's reputational score. Confirmed and upvoted facts will lead to both material and reputational gains for the author.

Posting own facts

Once a participant has created a verified account, they may create posts in all three categories. We expect to regulate the frequency of the posts: with more trust gained from peers, a participant may post more often. A fact posted should consist of a statement, justification, and the evidence. One has to select (or use the default) amount of cryptocurrency to back the statement. The author also has to select the time horizon for the discussion.

Contesting facts

Other experts may support or contest the posted fact. One may respond to the message and attach one's own evidence related to the fact. Note, by posting one agrees to the time set by the originator.

Each expert could set any number of crypto coins to back their own publication of a fact. After the time horizon comes to expiration, the fact's correctness is automatically evaluated based on the weighted opinion of the experts. Those who originally supported this fact will get the backing money of those who opposed it.

Placing correct information and supporting correct facts are incentivized by improving the trust and expert score for the participants. This will result in their higher weights for the next decisions. Note, one cannot contest one's own fact or downvote it. Currently, the project uses an uploaded copy of Wikipedia on a blockchain as its feed information and reference source.

Publishing an opinion

One may also publish an opinion. It is not possible to contest it, but it is allowed to comment, upvote, or downvote this publication. The comments cannot be deleted; they stay forever.

Publishing a belief

One can also publish a belief, which is classified in a broad sense as a religious view. One cannot comment or downvote it. However, one may contest the choice of the classification, and if there are sufficient opposing views, it can be re-classified to an opinion. We are planning to employ an AI system, which will help to classify and monitor the classification of the publications. Misclassification will result in a reduction of the trust score.

Reputation scoring

In addition to direct material incentives for publishing trustworthy information, a reputation scoring system serves as another layer of motivation. The prototype system is currently based on the EigenTrust methodology; however, we have designed and tested a more advanced proprietary algorithm that is more stable than EigenTrust against common reputation attacks.

Votes regarding specific sources or pieces of information will be visible to all nodes within the blockchain network, making data about integrity and fraud transparent to everyone. The aggregation of all votes will be weighted based on the credibility of each individual voter.

Thanks to the decentralized nature of the proposed blockchain solution, the data remains resistant to falsification or manipulation. This ensures that the information stays credible, tamper-proof, and efficiently managed through distributed storage. The architectural design of the solution also makes it impervious to control or shutdown by authoritarian regimes, as well as resistant to attacks involving anonymous, multiple accounts (Sybil-type attacks), self-promoting, and re-entry (or whitewashing) attacks.

Once the published position is cleared, the reputation of the originator will be updated. We are planning to limit the frequency of monetary transactions to prevent spamming and manipulation of the reputation scoring. Note that contesting statements made by accounts with a higher reputation score will be more rewarding than contesting accounts with low reputation scoring.

A new profession: fact bidders and fact checkers

This platform could give rise to a new profession: the professional fact-bidder, who profits from identifying inaccuracies in information while aiming for a fact-based, research-backed consensus.

Thanks to blockchain technology, the outcomes of contests can be traced even after they have expired and been executed.

These professional fact-checkers are incentivized to not only contest stories but also to publish their fact-checking research, thereby maximizing both their reputational score and profits. They have a strong motive to quickly and efficiently find and document accurate data. Moreover, fact-checkers can provide

verified external links to data sources, such as governmental and other institutional organizations, thereby enabling additional bids related to newly presented information.

Immutability of the Verified Information

Once one publishes a fact and it is verified, one can use a unique link to refer to it stored on the blockchain. We plan to create an addon to view verified links using regular browsers. Any published and verified fact, video, or picture cannot be faked since they are encoded on the blockchain, and a browser's addon will verify the hash sum for each link viewed.

Blockchain technology allows us to verify if a story/news/video/picture has not been altered, making it impossible to create edited copies of the verified and upvoted news (in particular, videos using AI technologies, like deep fakes) and distribute them like unedited information.

These copies can be verified in a decentralized manner by network participants, whose votes will be aggregated based on their trust and expertise. The collective opinion of experts with high-level reputations and regular participants is a very efficient tool to aggregate the available knowledge and translate it in a democratic, efficient, and unbiased way. Once new information becomes available, people are incentivized to share it in order to overturn the old, incorrect opinion and get a reward for this.

Our blockchain platform offers a transparent solution when it comes to the content of the news. All participants of the blockchain can trace the history of the communication, associated metadata, and all modifications and updates performed by the owner of this communication. The owner will also have an option to indicate their confidence in the news, which can be contested by fact-checkers.

The system allows news/information consumers to see if the originator has updated the initial unverified news, or has modified any content. This immutability applies to texts, pictures, audios, videos, and other data formats. Based on this data, the blockchain participants can evaluate their confidence in the trustworthiness of the news. The users may also set a trustworthiness threshold for their news feed to view more reliable news through subscription. Newsmakers with bad reputation will eventually be ranked under most thresholds, making themselves self-isolated.

Types of Accounts

There are two types of accounts in the Satya ecosystem: free and subscription-based. Free accounts on the platform allow users to consume information tagged as free; these accounts do not have publishing, voting/contesting or commenting rights. All unverified information is freely accessible. However, information that has undergone verification may not be accessible without a subscription, the decision is up to the owner of the verified information.

A non-free account, acquired through a reasonable subscription fee (about $1 a year), grants users the ability to publish news and contest the information posted by other participants. These non-free accounts require the owner's true identity to be verified. This measure reduces the likelihood of various types of Sybil attacks, as maintaining multiple digitally verified, paid accounts would be cost-prohibitive. Furthermore, while the platform does not take responsibility for any hate speech or criminal behavior by its members, such activities can readily be addressed by law enforcement.

Big news outlets have the option to register as a portal node, incorporating a hierarchical structure of sub-nodes. This setup could serve as an internal blockchain accessible only to employees of that portal.

Overall, the platform accommodates both free and subscription-based content distribution. While the cost of content could be nominal, referring to or distributing paid and verified content will automatically generate potential revenue for the content creator. This revenue becomes actualized only if the content is verified as truthful. Otherwise, the revenue is redirected to those who contest against the content's trustworthiness, serving as an additional deterrent against the propagation of fake news. Even minimally verified information carries some non-zero value, thereby creating an avenue for monetizing fact-checking efforts.

COMMUNICATION OUTLET

The platform will allow encrypted private communication between registered and digitally verified participants.

GROWTH POTENTIAL, METRICS AND REVENUE

TARGET MARKET:

We envision that our initial target market is communities of experts in politics, economy, and science, overlapping with the target audience of "Knowledge markets" like Answers.com, Baidu Knows, Quora, Reddit, Stack Exchange, etc. The central idea is to attract an initial cluster of known experts by providing them an easy opportunity— an initial invitation bonus in the platform’s cryptocurrency, with the option to exchange for USD after some agreed time period, conditioned on their contribution by publication and voting on published information.

OUR COMPETITORS:

Our competitors are social platforms such as Quora, Reddit, Telegram, and, as a decentralized platform, we will compete with other distributed platforms like Mastodon, DTube, etc.

Product or Service Validation:

We have developed an MVP. We have received a positive response from several testers. We are planning large-scale testing of the platform by inviting CUNY students. It has all the discussed basic functionality, allowing for account creation, information publication, voting, and backing the published facts with the internal currency. We still have to finalize the Self-Sovereign Identity and governmental ID verification part of the software.

Business Model:

The primary revenue will originate from the transaction fee (initially ~1%) for each settlement for published facts, plus subscription fees for premium accounts. The idea is to create initial momentum and support it by giving early adopters free tokens to participate in facts validation.

We expect that the project will become profitable with a sustainable volume of about 10,000 daily transactions.

Use of Funds:

We currently need additional funding to finalize the SSI (identification) component of our platform and hire security professionals to test the resilience of the platform to various types of cyberattacks. Some funds will be allocated for creating an advertisement campaign within the target communities of experts.

Legal and Regulatory Compliance:

There is a risk that the current business model may eventually require a gambling license. However, one may make a strong case that this is a knowledge/expertise remuneration framework.

Technological Foundation

The storage system is based on the proprietary blockchain, no gas for transactions. The frontend is designed using Google's Flutter (Dart language).

APPENDIX

We perform simulations how different proportion of various agents and vote aggregation schemes affect the outcome of the vote. Based on the simulations, one may predict the necessary number of trustable experts in the system to prevent high jacking and manipulation of the voting system. Voting is a central methodology for eliciting and combining agents' preferences and knowledge. Our methodology is based on the expert weight determination and update, which is a critical issue in the evaluation process and stability of the system. There are numerous voting rules and different voting systems which we are planning to test. There are at least several types of participants in the voting game. The types can be characterized by the honesty parameter (honest/dishonest) and knowledge (more knowledgeable /less knowledgeable).

Agent Variables:

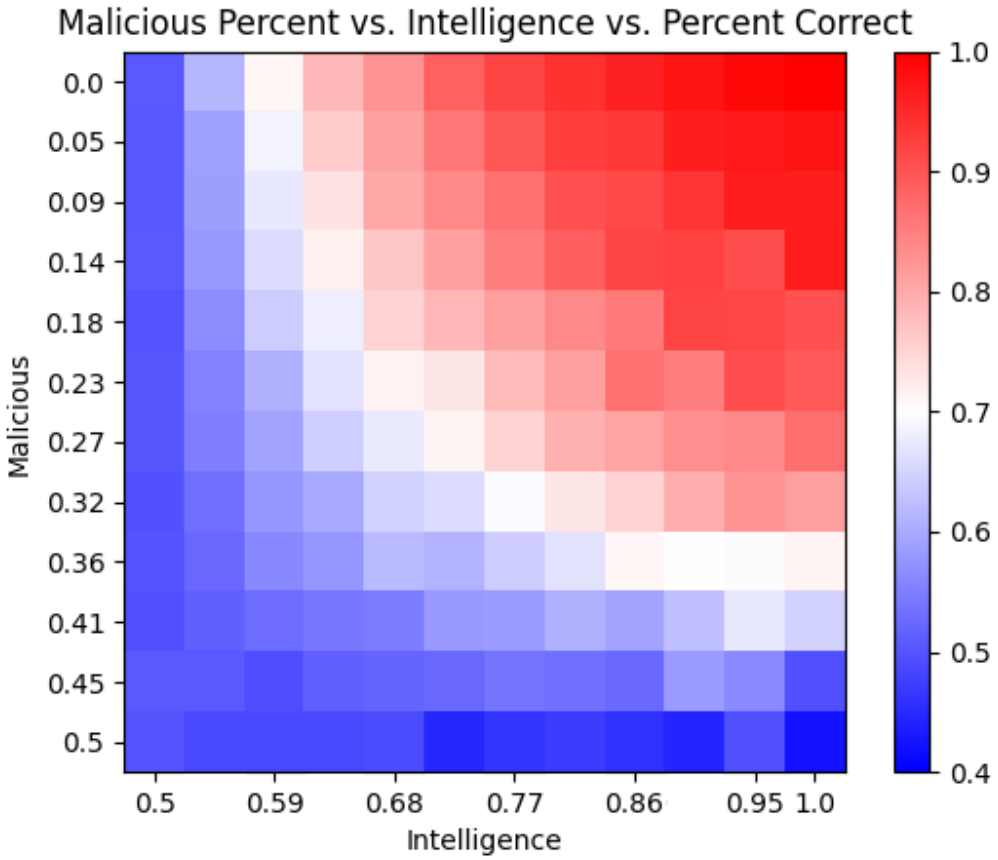
- **Intelligence**: How likely the agent is to know whether a fact is true or false
  - Intelligence 1 = Always knows whether true or false
  - Intelligence .5 = 50/50 (minimum possible intelligence used)
- **Agent Type**: The way the agent chooses its bet
  - **Belief**:
    - Votes fact is true or false depending on if agent believes fact is true or false (based on the agent's intelligence)
  - **Malicious**:
    - Votes opposite of what agent believes
  - **Agree Smarter**:
    - Votes true if creator agent has higher trust than own trust, else votes whatever it believes
  - **Highest Trust**:
    - Votes the most common vote among the agents with highest trust.
  - **Agree**:
    - Always votes true
  - **Disagree**:
    - Always votes false

Graphs:

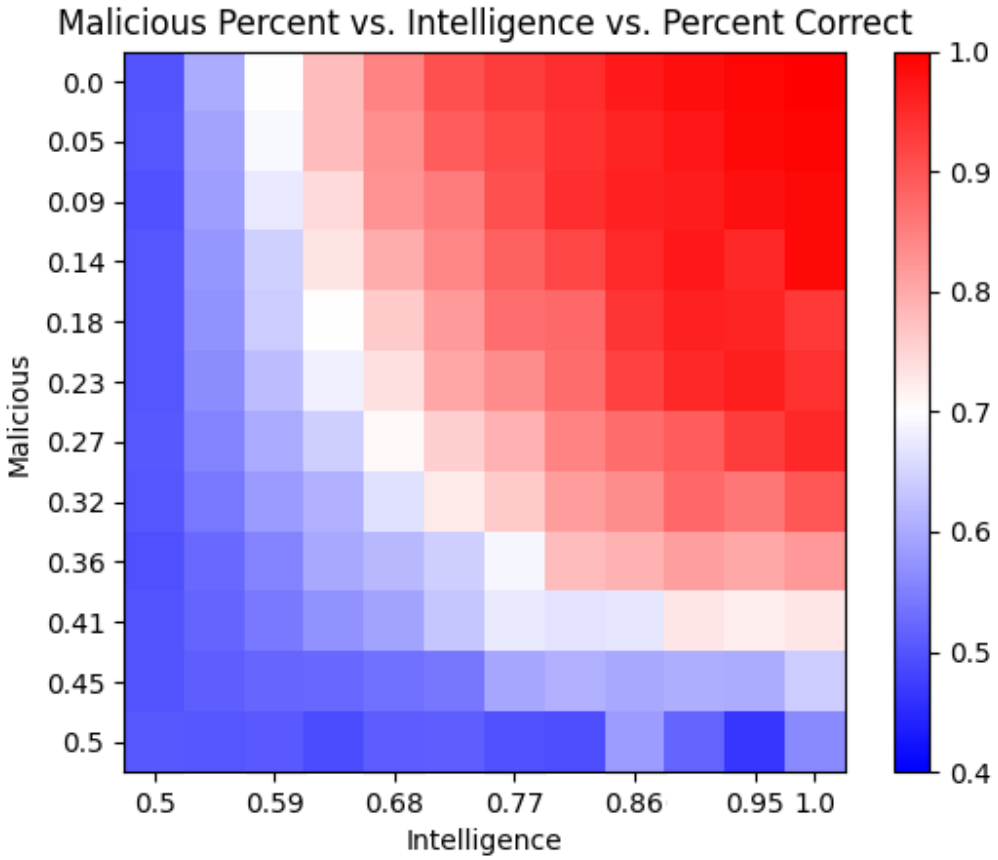
- **Only Belief and Malicious Agents**:

  Belief: 100%-50%
  Malicious: 0%-50%
  Agree Smarter: 0%
  Highest Trust: 0%
  Agree: 0%
  Disagree: 0%

  - Equal Trust (every agent has the same trust)



- Simple Trust (trust depends on number of other agent's that have agreed with it in the past)



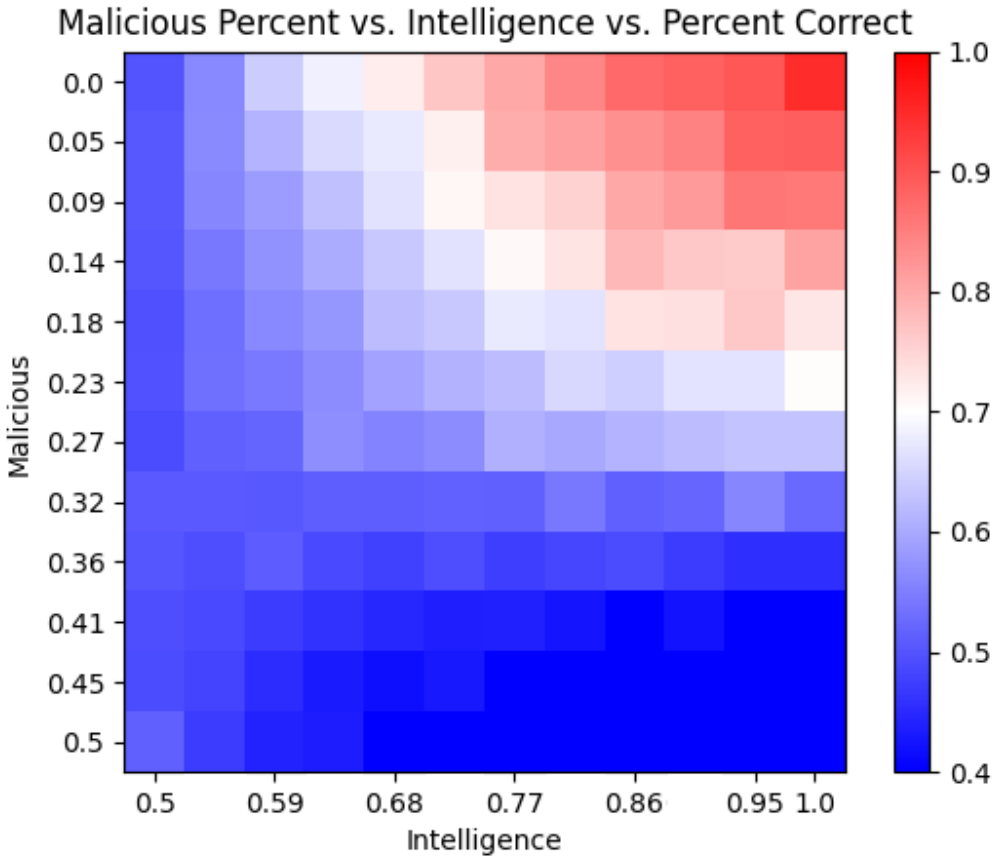
Notice how with the simple trust algorithm, the red region stretches further down showing that using trust algorithms makes games more resistant to malicious actors for all graphs

- **Equal Amount of Non Malicious Agents**:

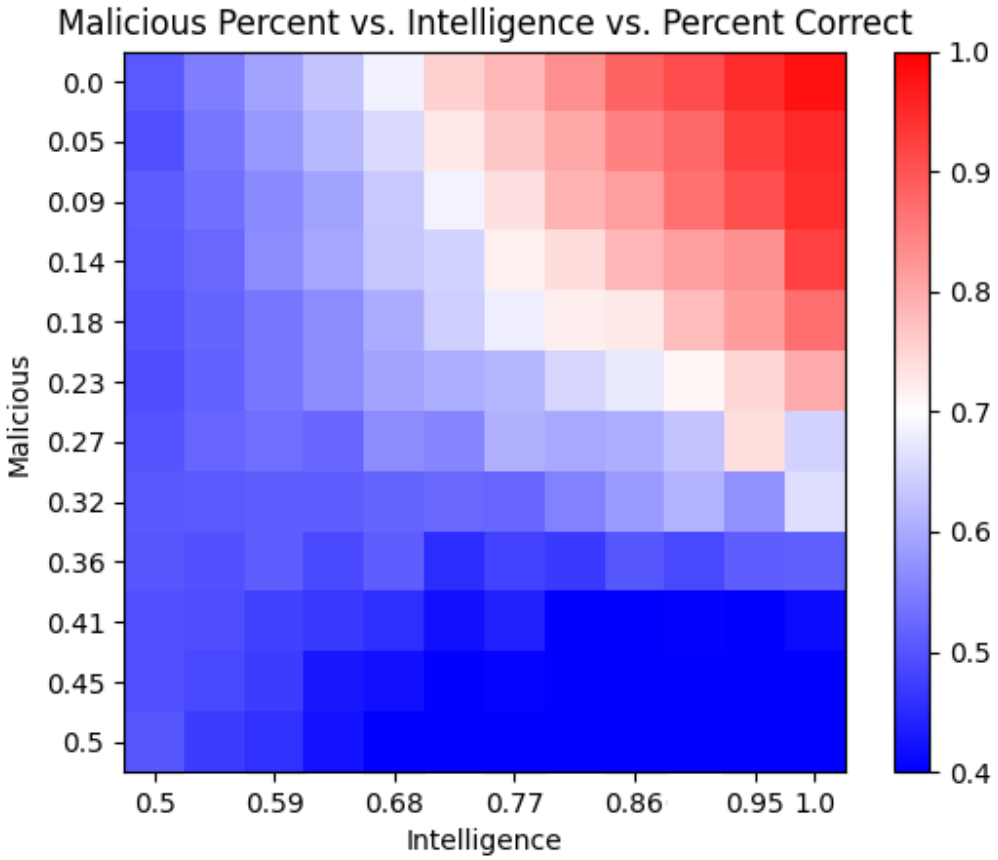
We split non-malicious agents into 5 equally populated groups

Belief: 20%-10%
Malicious: 0%-50%
Agree Smarter: 20%-10%
Highest Trust: 20%-10%
Agree: 20%-10%
Disagree: 20%-10%

- Equal Trust (every agent has the same trust)



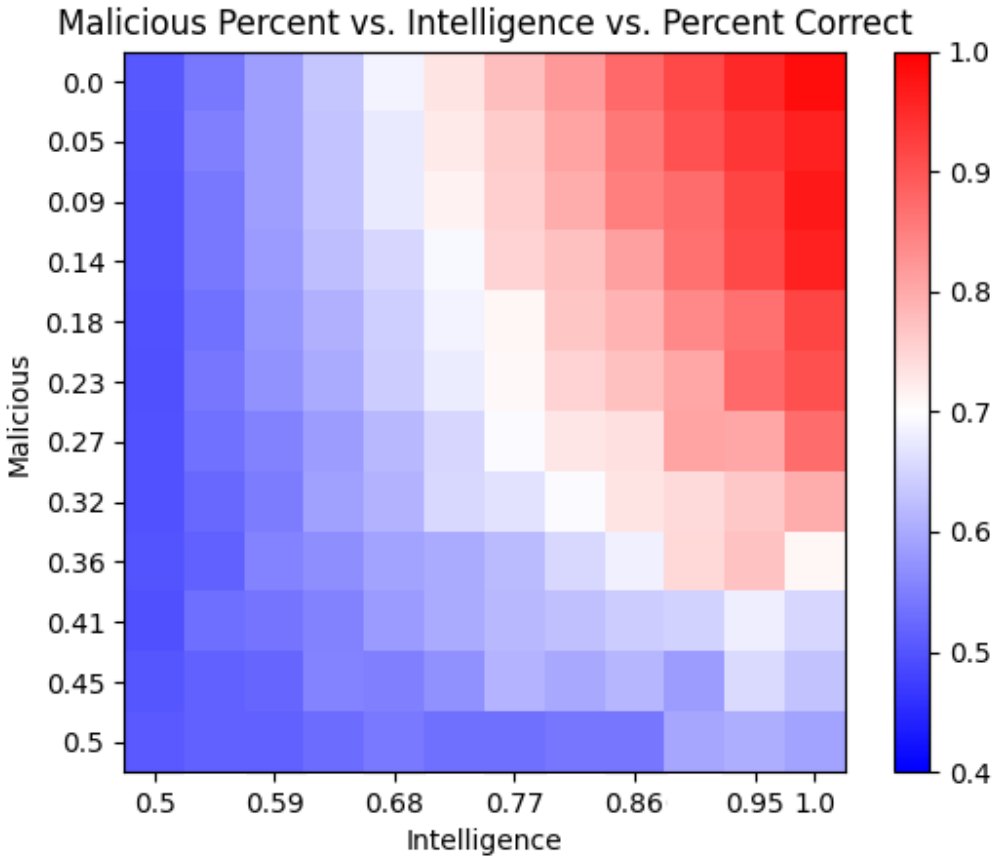
- Simple Trust (trust depends on number of other agent's that have agreed with it in the past)



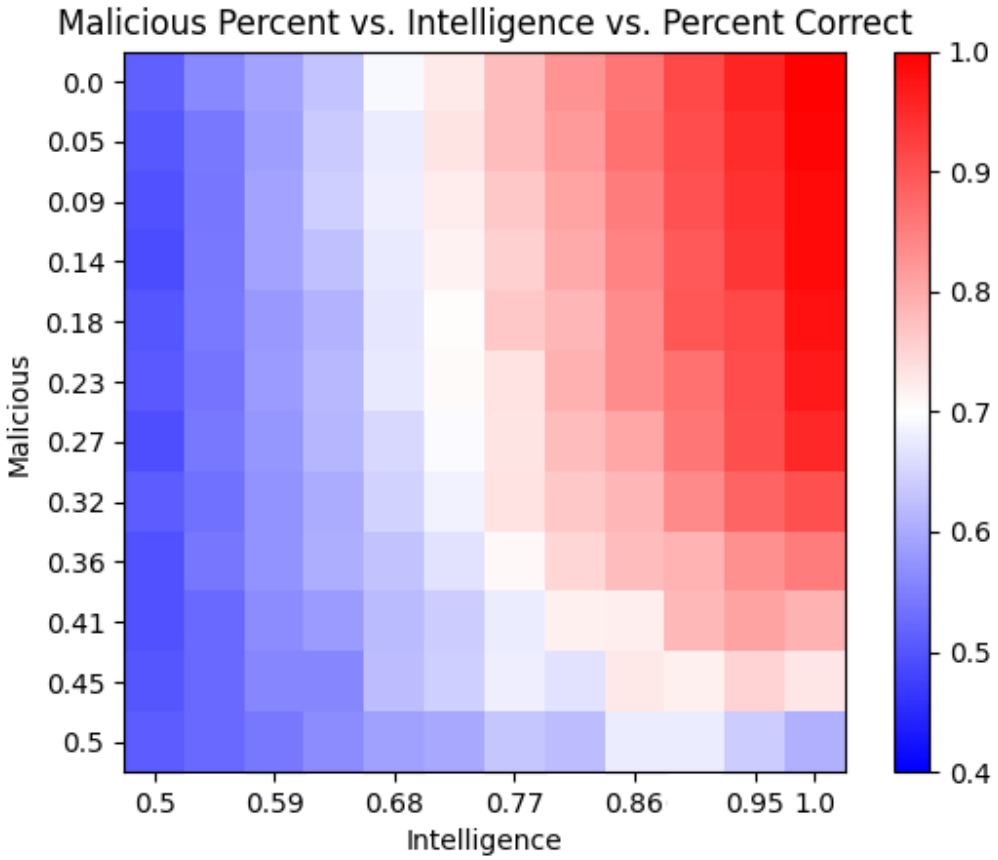
- **Alternative Ratios of Agents**:
  The ratio of belief to agree smarter to highest trust to agree to disagree stays constant while percentage of malicious changes.

  Belief: 22.2%-14.3%
  Malicious: 0%-50%
  Agree Smarter: 5.5%-3.6%
  Highest Trust: 66.6%-42.9%
  Agree: 2.7%-1.8%
  Disagree: 2.7%-1.8%

  - Equal Trust (every agent has the same trust)



- Simple Trust


Malicious Percent vs. Intelligence vs. Percent Correct
Malicious
0.0
0.05
0.09
0.14
0.18
0.23
0.27
0.32
0.36
0.41
0.45
0.5
0.5
0.59
0.68
0.77
0.86
0.95
1.0
Intelligence
1.0
0.9
0.8
0.7
0.6
0.5
0.4