9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Monetization of social media engagements increases the sharing of false (and other) news but penalization moderates it

Keywords: Decentralized Social Media, Misinformation, Polarization, Content Moderation, Content Monetization

Extended Abstract

Major social media companies are enacting to Decentralized Social Media (DeSo) and incorporate similar features. Reddit introduced 'Community Points', which are a measure of reputation to reward users for their activity and quality content. The points can be converted to 'Reddit Coins' and users can spend them only on Reddit. However, there are speculations about Reddit's plan to eventually turn the points to an Ethereum-based token, which would enable the owners to spend it anywhere.[1]. Twitter introduced 'Tips', which allows users to support content by sending fiat money, Bitcoin, or Ethereum to creators. More recently, on the 5th of November 2022, Elon Musk announced in a tweet that "creator monetization for all forms of content" will soon be added to Twitter.[2].

While the idea of DeSo is not new, the possibility of major platforms embracing monetary incentives of DeSo is both promising and worrisome. On the one hand, it might help to alleviate problems linked to content moderation, data protection, advertisement pricing, and compensating content developers [4]. On the other hand, while existing research shows that accuracy nudging decreases the probability of sharing fake news [2], and that monetary incentives may serve as such nudges [3], we know little about the effects of monetary incentives on the willingness to share different kinds of content. Hence, insights about their potential negative consequences is urgently required. To this end, we propose a framework (Fig 1) to understand how DeSo may affect user behavior and conduct 1) an online survey to explore how well these new financial incentives of DeSo are understood by people, and 2) an online survey experiment to examine how the monetary incentives of DeSo might affect the likelihood of posting different kinds of content.

We preregistered a target sample of 1,000 representative US participants recruited from Prolific. 1,067 participants started the survey and 1,022 of them completed it. The final sample (mean age = 45.40) included 517 women, 496 men, and 9 other gender identities. Respondents were presented with five statements about cryptocurrencies of which two were correct and three incorrect which they had to rate as either 'right' or 'wrong'. We observe that there is little variation in correct response rates. On average, respondents in the reference groups answered about 46% of the questions correctly. The only significant differences in a negative direction are found for older respondents (65+, relative to 18-24 years old) and —strikingly— the self-indicated most knowledgeable. More specifically, older respondents and respondents who reported to be "extremely knowledgeable" about crypto relative to those reporting to be "not at all knowledgeable" (Fig 2) had a poorer understanding of crypto.

[1] https://cointelegraph.com/news/reddit-to-reportedly-tokenize-karma-points-and-onboard-500m-new-users
[2] https://www.businessinsider.com/elon-musk-twitter-monetization-model-forms-content-youtube-2022-11

To provide evidence on the possible effects of the monetary incentives of DeSo on users' willingness to share news headlines, we conducted an online survey experiment with $n = 1,500$ participants from a nationally representative U.S. sample recruited from Prolific. The participants were presented with five true neutral and hyper-partisan and five misinformation news headlines about COVID19, which were randomly selected from [1], and were then asked about their willingness to share them. Participants in the treatment group were shown a set of statements about a hypothetical world in which platforms reward their users with digital asset for their *reputation points* calculated based on their user engagements. While in all treatments participants earn a fixed reward, the treatments vary across two dimensions: a) penalization for misinformation or hate speech, and b) explicit or vague calculation of points.

Fig 3 shows the effects of rewards, penalties, and calculation methods of points on the reported willingness to share news. We highlight four sets of results. First, in the absence of penalties, rewards for user engagements increase the willingness to share all kinds of news. Respondents are about 9% more likely to share misinformation when engagement is monetized ($coef = 0.089, se = 0.015, p = 0.000$). Second, the same effect holds if respondents can be penalized for sharing misinformation, although effect size decreases to 4% ($coef = 0.041, se = 0.016, p = 0.03$). Third, given the presence of rewards, respondents that cannot be penalized for sharing misinformation are about four percentage points more likely to share misinformation than those who can be subject to penalties ($coef = 0.043, se = 0.016, p = 0.02$). Fourth, the clarity of the link between rewards and user engagements, and between penalties and sharing behavior, has no effect on the willingness to share content of any kind.

Our main conclusion is that rewarding users monetarily for their user engagements increase the news sharing intention of them including misinformation. However, if monetary penalties are set in place, that would disincentivize sharing intention of misinformation, but does not diminish it. Our findings that monetary incentive outweighs monetary penalization and participants show higher sharing intention of hyperpartisan news when misinformation is penalized are in agreement with the idea that people engage disproportionately higher with "borderline content" [5]. The decentralization of social media may further increase the misinformation problems that social media already face.

References

[1] Pennycook, G., J. Binnendyk, C. Newton, and D. G. Rand (2021, July). A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra: Psychology 7*(1), 25293.

[2] Pennycook, G., Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand (2021, April). Shifting attention to accuracy can reduce misinformation online. *Nature 592*(7855), 590–595.

[3] Prior, M., G. Sood, and K. Khanna (2015, December). You Cannot be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions. *Quarterly Journal of Political Science 10*(4), 489–518.

[4] Sockin, M. and W. Xiong (2022). Decentralization through tokenization. Technical report, National Bureau of Economic Research.

[5] Zuckerberg, M. (2018). A blueprint for content governance and enforcement. *Retrieved from Facebook Newsroom website: https://www. facebook. com/notes/mark-zuckerberg/ablueprint-for-content-governance-andenforcement/10156443129621634*.

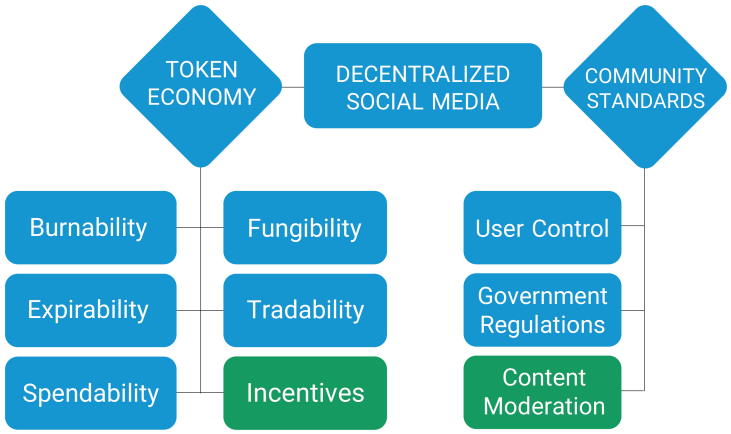

Figure 1: Main features of a decentralized social media that may affect the posting activity of users. Green boxes denote the parts studied in this paper. Burnability indicates whether a token can burned to terminate a right or revoke access. Expirability refers to whether a token can be expired after some time. Spendability indicates whether a token can be used to gain access to services or pay fees. Fungibility says if a token is interchangeable with other tokens. Tradability illustrates if a token can change ownership within a platform or on secondary markets. Finally, incentives schemes can be divided to two categories: 1) incentive enablers, which refer to what token holder can do with the token, and 2) incentive drivers, that indicate why a token holder engages with incentivized behaviour (e.g. gain reward or reputation).

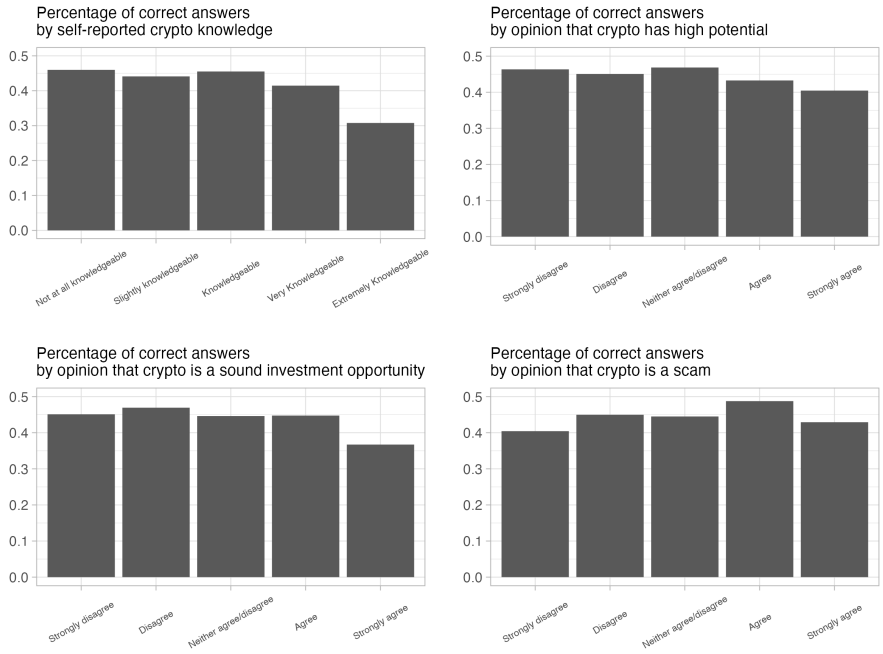

Figure 2: Fraction of participants who answered 5 questions about cryptocurrency correctly by their self-reported knowledge and opinion about cryptocurrency. In study 1, n = 1,022 American individuals from Prolific were presented with five statements about cryptocurrencies of which two were correct and three incorrect which they had to rate as either as 'right' or 'wrong'. In general, the results show that the more participants think they know about cryptocurrency or have more positive opinion about it, the less likely they answer correctly to questions. (top left) Participants who reported to be 'extremely knowledgeable' about cryptocurrency are less likely to answer correctly to questions. (top right) Participants who reported to be 'strongly agree' with the high potential of cryprocurency are less likely to answer correctly to questions. (bottom left) Participants who reported to be 'strongly agree' that cryptocurrency is a sound investment opportunity are less likely to answer correctly to questions. (bottom right) Participants who reported to be 'strongly disagree' that cryptocurrency is a scam are less likely to answer correctly to questions.

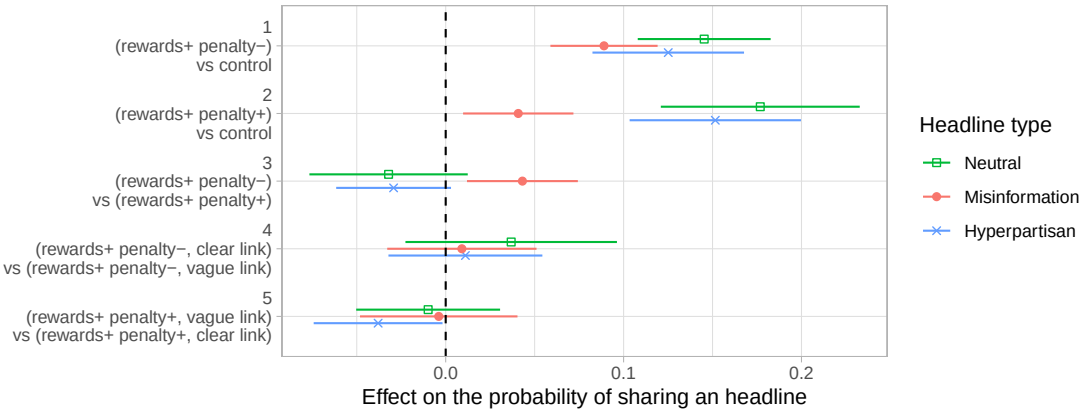

Figure 3: Point estimates from OLS regressions with 95% CI. Effect of monetary rewards for user engagements and penalties for problematic sharing behavior, as well as of the clarity of the link between rewards/penalties and user engagements/behavior, on the willingness to share neutral and hyperpartisan news, and misinformation. Without penalties, rewards increase the sharing intention of misinformation by about nine percentage points, compared to the control group. With penalties, the effect decreases to about four percentage points. The clarity of the link between rewards/penalties and user engagements/behavior has no effect.