1 Project Eiden

Problem. The main goal of this project is to reduce productive inefficiencies in science by promoting more open exchange of research questions, ideas, and data. The generation of research ideas has some random component, and research teams do not necessarily have comparative advantages to develop their own ideas and questions. For example, Watson and Crick (Cambridge) based their DNA model on yet unpublished X-ray photographs from the research group at King's College. While in this story both groups eventually ended up with Nobel prizes, in general the exchange of early-stage research ideas between teams is limited due to fears of scooping. As a result, many potential projects do not happen. Researchers can also donate ideas they do not want to pursue, but there is no mechanism or accepted practice to attribute ideas and without attribution the personal returns of donations might be too small.

The project would also help to reduce two other known inefficiencies. First, a large part of scientific research output is partially or completely redundant (Azoulay et al., 2018). This happens because scientists know where the research frontier is but have little knowledge about competing projects. Additionally, scientists tend to do more redundant research than socially optimal because they do not internalize the effects of their research on competitors (Dasgupta and Maskin, 1987).

Approach. The project aims to create an open science repository with an incentive-compatible peer-review mechanism[1] The platform would have an interactive database of research projects at all stages of work from a research idea to a final paper with several key features:

1. Publishing research projects on stages of idea/proposal/working paper.
2. Requesting and publishing referee reports (including reports anonymous for the author and/or third parties).
3. Publishing less formal or detailed comments.
4. Rating research projects, referee reports and comments.
5. Hosting grant competitions.

The new platform aims to provide sufficient participation incentives for scientists at different career stages. Highly established scientists can use it to increase their footprint in science and

[1]Simply speaking, the mechanism is incentive-compatible if agents get higher payoffs by acting truthfully. For example, asking a child which dessert they want to pick in a cafe is incentive-compatible because a child gets the highest payoff by answering truthfully. This definition encompasses strategyproof mechanisms and Bayesian Nash equilibrium incentive-compatible. In strategyproof mechanisms agents always receive higher or equal payoffs by reporting truthfully regardless of choices of others. A Bayesian Nash equilibrium incentive-compatible mechanism has at least one equilibrium in which acting truthfully gives the best expected payoff conditional on others acting truthfully.

advertise their research groups. Early career scientists would benefit by learning through the feedback and by increasing their professional exposure.

As an open science repository, it would compete with several other platforms, including ResearchGate, Academia.edu, Mendeley, arXiv and SSRN. While these platforms mostly work successfully as preprint repositories, they do not have formal peer-review systems and do not seem to work well to get accurate feedback on early research. They also do not upload research ideas and so provide limited opportunities to direct research and seek collaborators.

Reputation Metric. As some other open science platforms, the project would use reputation metrics to reward users for contributions to the community such as sharing research projects, writing referee reports and giving accurate peer evaluations. In contrast to existing platforms, we specifically design the reputation metric mechanism to reward truthful peer evaluations and hence make the reputation an informative signal of researcher's quality.

The public reputation metric would determine visibility of users' projects and impacts of their peer evaluations. The platform will have an incentive-compatible peer review system for two types of user's submissions: research projects and referee reports. Each user in the system would have an ability to rate all the projects and all the referee reports of other users[2]. Users will be able also to like comments provided to projects and referee reports, but these likes would not affect neither commentator nor rater's reputations.

We expect the reputation metric to proxy for the quality of user's research output. Reputation of a user i is a sum of their initial reputation $R_i(0)$ (such as academic rank and H-index), reputation of the user's projects $R_i^p(t)$, user's referee reports $R_i^r(t)$ and accuracy score of peer evaluations $E_i(t)$ given by user i for all the projects and referee reports they reviewed. The accuracy score of peer evaluations $E_i(t)$ for user i is the sum $E_i^p(t)$ of the accuracy scores for all the research projects rated by i plus the sum $E_i^r(t)$ of accuracy scores for all the referee reports rated by the user. We can describe the reputation by the following formula if abstracting from weighting of different components:

$$R_i(t) = R_i(0) + R_i^p(t) + R_i^r(t) + E_i(t), E_i(t) = E_i^p(t) + E_i^r(t)$$

Why would users value their reputations? The reputation would be valuable if it becomes an informative signal of user's quality as a researcher and can affect their visibility and their careers[3]. Assume that we have some imperfect real-world measure of researcher's professional success X_i such as their Hirsch index, citation count or an academic ranking of the employer. The reputation metric is informative and valuable if it improves the expectation of future

[2]Subject to potential limitations on the total impact or number of ratings to limit the power of overactive users.

[3]Users would not be able to observe the composition of others' reputations. If users decide that the reputation of research projects is much more informative than the rest of the metric, they would lose incentives for accurate peer evaluations. This would, in turn, erode the quality of peer evaluations and can unravel the rest of the reputation metric.

success conditional on other publicly available information[4]:

$$\frac{\partial E(X_i(t+T)|X_i(t), R_i(t))}{\partial R_i(t)} > 0, T > 0, \forall t,$$

An informative signal would still have zero value if its users do not know that it is informative. The sufficient condition is to require that it is common knowledge that the ratings affecting the reputation are (largely) truthful. This makes it extremely important to rely on incentive-compatible accuracy ratings and to make sure that the platform's activity stays in truthful equilibria.

Accuracy Scores. Accuracy scores reward users to provide truthful peer evaluations of research projects and referee reports. Users providing more accurate ratings expect higher increases in their reputation. In contrast, providing distorted or noisy ratings harms the reputation score.

The project would use different approaches to calculate accuracy scores of peer evaluations for research projects and for referee reports. There are multiple though imperfect ways to observe whether the research project turns out to be good: publication success and impact factor of the journal, citations count or the number of downloads. In contrast, we have less options to verify quality for referee reports except evaluations by editors and authors. It necessitates the use of different mechanisms with the research projects relying on a more robust approach based on proper scoring rules.

The platform would use the quadratic score function (Selten, 1998) to measure the accuracy of beliefs about research projects. This is the proper score function mapping on the reported probabilistic belief about an event and the event's actual realization to one number. By definition of the proper scoring function, its expected value is the highest when the true probability is used. Hence the mechanism is incentive-compatible in the strictest sense as long as users are risk-neutral (care only about the expected value).

In order to get a project rated, the author posting it on the platform has to specify one or more publicly verifiable binary measure of the project's success. As an example, the author can ask if paper(s) resulting from the project would be cited at least 10 times within three years after the project's progression to the final stage. The research project's rating is an estimate of the probability for that event. If user i predicts that the probability of this event is p_{ik} then the accuracy score of this rating will be:

$$e_{ik} = S(p_{ik}, o_k) - S(p_{Bk}, o_k))$$

$$S(p, o_k) = 1 - 2(o_k - p)^2$$

Where o_k is the indicator of the event:$o_k = 1$ if it occurs (e.g., the papers get cited at least

[4]The signal is still informative even if its predictive power comes solely from self-fulfilling expectations.

10 times), $o_k = 0$ otherwise. The component $S(p_{Bk}, o_k) = 1 - 2(o_k - p_{Bk})^2$ represents the benchmark score to rescale the accuracy. The value p_{Bk} is the benchmark prediction we obtain by using either the average community prediction or the algorithmic prediction. Algorithmic prediction would use basic machine learning techniques to estimate the probability based on verifiable project's characteristics such as completion stage, word count, and reputation of its authors. It is crucial that the benchmark score does not depend on user i's actions and hence does not affect change their optimal strategy conditional on participation.

If the benchmark score comes from the AI, users improve their reputation by predicting better than the algorithm. In contrast to the community average prediction, AI rating can be made public almost immediately and before any private ratings. Users will be able to evaluate potential reputation gains by comparing their private beliefs against the AI rating. It should motivate users to seek underrated/overrated projects and incorporate new private information into their predictions. The simplest AI prediction involves just a constant score for all the items within the category (projects/reports).

Quadratic score function is an incentive-compatible mechanism to elicit beliefs under two plausible assumptions (Selten, 1998). First, we need to assume that users prefers more to less reputation. Second, we need to assume that users are risk-neutral or, in other words, they care only about their expected reputation and not about other moments of its potential distribution. As we argue before, the reputation should be valuable for users if it is an informative signal of their quality as researchers and if users value how others perceive their quality[5]. There is no reason to know in advance if the users are going to be risk-averse to reputation, but the bias from risk-averse users is unlikely to make reported ratings uninformative. Both theory and laboratory experiments show that risk-averse raters bias their reported beliefs by over-reporting low beliefs and under-reporting high beliefs (Armantier and Treich, 2013; Andersen, Fountain, Harrison and Rutström, 2014). This bias does not change the strictly monotonic relationship between beliefs and reports and can be ex-post corrected (Offerman et al., 2009).

Several existing platforms already use proper scoring rules or other reputation-based prediction markets to measure beliefs of their users. Metaculus allows users to bet their reputation on outcomes of future events by using the combination of the log-score and other proper scoring functions. On 30 March 2021 Metaculus had almost 16 thousand registered users (forecasters)[6]. Their internal report show that median reported beliefs (community beliefs) from the start at 2016 to February 2021 had a Brier proper score of 0.122 meaning significant gain over the random reporting[7]. Hollywood Stock Exchange platform existing since 1996 has several surprisingly accurate predictions of box-opening takes for movies (Mann, 2016).

Because there is no easily observable ground truth for referee reports, their

[5]This assumption may not hold for people expecting to leave academic community in near future (retirement, death, grave ethical violations).

[6]Based on its API: https://www.metaculus.com/api2/users/

[7]Unfortunately, there is no published comparison of forecasts between prediction markets with real money and Metaculus.

truthful ratings would rely on the approach of Radanovic et al. (2016). In theory, the approach elicits hidden correlated information from multiple agents in crowd-sourcing applications, when getting this information requires some effort and no verification is available to the center. For this platform, hidden information is the quality of the referee report. A user can rate a referee report k either as "good" (g) or "bad" (b). We need at least two raters for each report to calculate the accuracy score, because it depends on the match between the user's i rating and the rating of some randomly chosen other user (peer) p. Let x_{ik} denote the rating of user i and x_{pk} as the rating of some peer p. Then the accuracy score e_{ik} of user i for reviewing the report k is:

$$e_{ik} = \alpha(\tau(x_i, x_p) - 1)$$

$$\tau(x_i, x_p)e_{ik} = \begin{cases} \frac{1}{R(x_i)}, x_i = x_p \\ 0, x_i \neq x_p \end{cases}$$

Here $R(x_i)$ measures the proportion of ratings equal to x_i for similar tasks. In the original mechanism, this proportion is equal to the proportion of total tasks which are assumed to be similar. Referee reports appear to differ even before reading them. For this reason, the number $R(x_i)$ is made to be dependent on the report's visible characteristics.

Radanovic et al. (2016) show that this mechanism is incentive-compatible: telling the truth gives the highest expected accuracy score if others also tell the truth[8]. Users would benefit from giving truthful evaluations, because in equilibrium the expected accuracy scores are strictly positive. The mechanism is also "immune" to uninformative equilibria in which users submit the same rating for all the item (e.g., all referee reports are "good"). Any uninformative equilibria in this mechanism produces exactly zero expected payoffs/accuracy scores, and users lose interest in submitting uninformative ratings.

[8]They show that thuthful reporting is a subjective equilibrium of this mechanism meaning that it gives the highest payoff for any admissible belief about priors of other users.

References

Andersen, Steffen, John Fountain, Glenn W. Harrison, and E. Elisabet Rutström (2014) "Estimating subjective probabilities," *Journal of Risk and Uncertainty*, 48 (3), 207–229, 10.1007/s11166-014-9194-z.

Armantier, Olivier and Nicolas Treich (2013) "Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging," *European Economic Review*, 62, 17–40, 10.1016/j.euroecorev.2013.03.008.

Azoulay, Pierre et al. (2018) "Toward a more scientific science," *Science*, 361 (6408), 1194–1197, 10.1126/science.aav2484, Publisher: American Association for the Advancement of Science Section: Policy Forum.

Dasgupta, Partha and Eric Maskin (1987) "The Simple Economics of Research Portfolios," *The Economic Journal*, 97 (387), 581–595, 10.2307/2232925, Publisher: [Royal Economic Society, Wiley].

Mann, Adam (2016) "The power of prediction markets," *Nature News*, 538 (7625), 308, 10.1038/538308a, Section: News Feature.

Offerman, Theo, Joep Sonnemans, Gijs Van De Kuilen, and Peter P. Wakker (2009) "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes*," *The Review of Economic Studies*, 76 (4), 1461–1489, 10.1111/j.1467-937X.2009.00557.x.

Radanovic, Goran, Boi Faltings, and Radu Jurca (2016) "Incentives for Effort in Crowdsourcing Using the Peer Truth Serum," *ACM Transactions on Intelligent Systems and Technology*, 7 (4), 48:1–48:28, 10.1145/2856102.

Selten, Reinhard (1998) "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1 (1), 43–61, 10.1023/A:1009957816843.