NBER WORKING PAPER SERIES

WILL USER-CONTRIBUTED AI TRAINING DATA EAT ITS OWN TAIL?

Joshua S. Gans

Working Paper 32686
http://www.nber.org/papers/w32686

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2024

Joshua Gans has drawn on the findings of his research for both compensated speaking engagements and consulting engagements. He has written the books Prediction Machines, Power & Prediction, and Innovation + Equality on the economics of AI for which he receives royalties. He is also chief economist of the Creative Destruction Lab, a University of Toronto-based program that helps seed stage companies, from which he receives compensation. He conducts consulting on anti-trust and intellectual property matters. He also has equity and advisory relationships with a number of startup firms. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Will User-Contributed AI Training Data Eat Its Own Tail?
Joshua S. Gans
NBER Working Paper No. 32686
July 2024
JEL No. D70,H44,O31

ABSTRACT

This paper examines and finds that the answer is likely to be no. The environment examined starts with users who contribute based on their motives to create a public good. Their own actions determine the quality of that public good but also embed a free-rider problem. When AI is trained on that data, it can generate similar contributions to the public good. It is shown that this increases the incentive of human users to provide contributions that are more costly to supply. Thus, the overall quality of contributions from both AI and humans rises compared to human-only contributions. In situations where platform providers want to generate more contributions using explicit incentives, the rate of return on such incentives is shown to be lower in this environment.

Joshua S. Gans
Rotman School of Management
University of Toronto
105 St. George Street
Toronto ON M5S 3E6
and NBER
joshua.gans@rotman.utoronto.ca

1 Introduction

Recent advances in artificial intelligence (or AI) are all significant improvements in computational statistics, specifically, the statistics of prediction (Agrawal et al., 2018), through machine learning based on neural networks. AI prediction is powered by data. There is data that is used as an *input* to AI algorithms to generate specific predictions. That data comprises personal information for, say, recommendation engines or prompts for generative AI. However, data was also used to create the AI algorithms in the first place. Extending on the neural metaphor that underlies machine learning, such data is called *training* data. Such data is used to determine the parameters (or weights) in the AI algorithm.

This paper examines a specific challenge that may impact the creation of suitable training data that comes from users' efforts. Such effort might be used to classify existing data and identify suitable outputs to assist in supervised learning. However, contributors might also provide useful information that can be used directly as training data in, say, generative AI. The question is whether the use of such data in AI training might cause feedback that impacts the creation of such data in the future.

Consider data such as users' written materials, artistic choices, and answers to questions in conversations and on platforms can be used to train AI algorithms to generate predictions of outputs from natural language prompts. Perhaps nowhere has this been seen more clearly than in generative AI tools for computer coding. These tools are trained on a corpus of open-source and other code. Moreover, by using sources where users post their coding issues, requests or queries and other users provide responses to those posts, generative AI can more tightly predict the appropriate responses to coding-related queries.

One platform that gained initial attention is Stack Overflow. On Stack Overflow users post queries regarding computer code that contributors answer. The contributors are not paid, although they are rewarded with various badges that can be displayed publicly and a reputation score. The more valuable contributions are found to be, the greater the rewards. As of 2022, there were over 24 million answers to 35 million questions. These were all publicly available and were used as training data by generative AI providers. One of those, OpenAI's ChatGPT, was able to provide answers to queries that might have otherwise been posted to Stack Overflow. One study found that ChatGPT's launch in 2022 caused an estimated 16 percent reduction in queries posted to Stack Overflow (del Rio-Chanona et al., 2023). This raised concerns that the availability of training data for future AI may be constrained by the use of AI today.

> ChatGPT is such a good programmer that the savvy developers I know aren't using Stack Overflow anymore – and yet it's partly by studying Stack Overflow

> that ChatGPT became such a good programmer. ...
>
> Where will this process take us? Stack Overflow was special because it drew out practical know-how that had, till then, lived only in programmers' brains; it condensed and organized that knowledge so that everyone could see and benefit from it. Chatbots that slowly siphon traffic away from sites like Stack Overflow obviously threaten that process. ... (Somers, 2023)

The concern is that the use of user-contributed training data might give rise to a 'doom loop' where AI eats its own source. "[I]f LLMs like ChatGPT present substitute traditional ways of searching and interrogating the web, then they will displace the very human behavior that generated their original training data.... As people begin to use LLMs instead of online knowledge repositories to find information, contributions to these repositories will likely decrease, diminishing the quantity and quality of these digital public goods." (del Rio-Chanona et al., 2023)[1] The reduction in user contributions on Stack Overflow suggested there was some substitution, but there was also evidence that this effect was limited, as the answers given were possibly of higher quality and the questions were, on average, more complex (Gallea, 2023).

In this paper, I present a simple model to address the question of whether, in domains where training data come from user contributions, the launch and subsequent use of AI products that potentially reduce the demand for such contributions will lead to a dearth of training data upon which to continue to improve those AI products? The simple intuition expressed above is that AI can cause a situation where less data is available for training and less incentive to contribute such data. However, the question here is: When this is investigated in a model that describes what drives user contributions in the first place, is the outcome one where AI faces intrinsic future limitations? The answer, as we shall see, is that AI is unlikely to do harm in the ways that commentators are concerned about.

2 A Simple Model of User Contributions

The model here is inspired by the query-and-answer format of Stack Overflow. There are n contributors available. Users post queries (or questions), the answers for which have an expected value to all users of V. There is a cost to a contributor for supplying an answer. That cost is made up of a common cost, C, representing the broad difficulty of the problem and an idiosyncratic cost, c_i, which is the cost incurred by contributor i. Idiosyncratic costs

[1] There was also a residual concern that contributors may use ChatGPT to provide answers on Stack Overflow in order to game reputational rewards (Xu et al., 2023). Stack Overflow prohibited such AI use, which seemed to have an effect; see Borwankar and Khern-am nuai (2023).

are independently and identically distributed according to an atomless, $F(c_i)$ on $(0, \bar{c}] \subset \mathbb{R}^+$. It is assumed that $V \in (C, C + \bar{c})$. If an answer is provided, all contributors receive a payoff αV that reflects the value they place on answers being available to users in general. That is less than the value to a general user, i.e., α is assumed to be less than 1. In addition, if a contributor provides an answer themselves, they receive additional utility of u that may be intrinsic or based on career concerns or reputational signalling (Lerner and Tirole, 2002).[2] We focus on the interesting case where $u \leq C$.[3]

The timeline for the model is as follows:

1. A user posts a query to the platform.
2. One contributor is selected at random to answer the query. If the query is answered, that contributor receives a payoff of $\alpha V + u - (C + c_i)$ and other contributors receive payoffs of αV.
3. If a query is not answered, one period advances (at a discount factor of δ) and stage 2 is repeated with the selection of a new contributor.

The assumption that one contributor is selected at random can be interpreted as any contributor who sees a query believes they are the only person seeing it at that time (for instance, time periods could be arbitrarily short).[4]

As noted above, by providing an answer, a contributor with cost realisation c_i, earns an expected payoff of $\alpha V + u - (C + c_i)$. If i does not contribute, then the query remains, and someone else may answer it. This is where a potential free-rider effect arises as the contributor can still realise αV, albeit with some delay. To calculate the expected payoff, v_0 from not contributing, suppose that there exists a $\hat{c}$ such that all those contributors with $c_i \leq \hat{c}$ contribute while others do not. In that case,

$$v_0(\hat{c}) = F(\hat{c})\alpha V + (1 - F(\hat{c}))\delta v_0(\hat{c}) \Leftrightarrow v_0(\hat{c}) = \frac{F(\hat{c})}{1 - (1 - F(\hat{c}))\delta}\alpha V$$

We can find $\hat{c}$ by equating this with the payoff from contribution, that is:

$$\alpha V + u - (C + \hat{c}) = \delta \frac{F(\hat{c})}{1 - (1 - F(\hat{c}))\delta}\alpha V$$

[2]This type of signalling has been found to account for a significant portion of motivation on Stack Overflow (Xu et al., 2020), and for open source software contributions (El-Komboz and Goldbeck, 2024).

[3]u itself can be a design choice for a contribution platform; however, here we do not examine these considerations. See Ghosh and Hummel (2011), Ghosh and McAfee (2011), Ghosh (2013) and Easley and Ghosh (2016).

[4]Engers and Gans (1998) have a similar set up in their model of refereeing but with fixed delay costs and no explicit individual incentives to contribute.

This assumes that $\hat{c} \leq \bar{c}$ (if not, $\hat{c} = \bar{c}$). Note that $\hat{c}$ is increasing in αV and $u - C$ and decreasing in δ. Intuitively, a higher $\hat{c}$ implies that a random contributor is more likely to answer the query only when their personal payoff $u - C$ and common payoff, αV, is higher. The reason for the latter is that the contributor internalises the delay in public good generation, which is positively related to the level of the common payoff. Similarly, when δ is higher, the contributor becomes more patient and hence, *less* willing to contribute.

3 The Model with Artificial Intelligence

Suppose now that an AI product is available that is trained on the queries and their answers. In this situation, a user can query the AI rather than post a query on the platform and, if possible, receive an answer immediately. The user can evaluate whether the AI answer is suitable or not and, if not, post the query to the platform. It is supposed that it is costless for users to query the AI, and therefore, they will all do so before posting to the platform.

It is also assumed that the AI is more likely to be able to answer an easier question, implying that the least costly queries will not be posted on the platform. This selection effect does not change the distribution of idiosyncratic contributor costs but does change the probability that a contributor will be able to answer any given posted query. That is, if all problems for which total cost is less than $C + \underline{c}$ can be undertaken by the AI, then the probability that a random contributor has a cost that is higher than this but lower than some threshold $\hat{c}$ is $F(\hat{c}) - F(\underline{c})$. This is a lower probability than would occur without the AI product being available.

Given this, the following proposition can be demonstrated.

Proposition 1 *Suppose that the AI can answer (without cost and immediately) all queries with a cost less than $C + \underline{c}$ where $\underline{c} > 0$. In equilibrium, if given the opportunity to answer a problem given C, without AI, contributors with $c_i \leq \hat{c}$ provide answers, and, with AI, contributors with $c_i \leq \hat{c}_{AI}$ provide answers with $\hat{c}_{AI} > \hat{c}$. AI increases the realised social value for any given query.*

Proof. Assume, for the moment, that $\hat{c}_{AI} \in (\underline{c}, \bar{c}]$. The equilibrium condition for $\hat{c}_{AI}$ is

$$\alpha V + u - (C + \hat{c}_{AI}) = \delta \frac{F(\hat{c}_{AI}) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}) - F(\underline{c})))\delta} \alpha V$$

Note that the right-hand side of this equation is at $\hat{c} = \hat{c}_{AI}$ strictly lower than the right-hand side of the equilibrium condition for $\hat{c}$ which implies that $\hat{c}_{AI} > \hat{c}$. Given that $u \leq C$, the overall production of answers has a lower cost with AI. Turning to realised social value, note

that, for a given query of difficulty, C, with AI, the total expected social value is:

$$F(\underline{c})V + (F(\hat{c}_{AI}) - F(\underline{c}))\,V + (1 - F(\hat{c}_{AI}))\delta\frac{F(\hat{c}_{AI}) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}) - F(\underline{c})))\delta}V$$

which is greater than $F(\hat{c})V + (1 - F(\hat{c}))\delta\frac{F(\hat{c})}{1-(1-F(\hat{c}))}V$, the expected social value without AI.

Note that as $\alpha \leq 1$, by the assumption that $V \leq C + \bar{c}$, the threshold by which $\alpha V + u - C = \hat{c}_{AI}$ will necessarily imply that $\hat{c}_{AI} \leq \bar{c}$, as per the (holding) assumption made above.

Turning now to evaluate the other (holding) assumption, i.e., whether $\hat{c}_{AI} > \underline{c}$, suppose that $\hat{c}_{AI} \leq \underline{c}$. In this case, with AI, the right-hand side of the contribution equation is 0, which implies that a contributor will set its threshold, $\hat{c}_{AI}$ so that $\alpha V + u - C = \hat{c}_{AI}$. If $\hat{c}_{AI} \leq \bar{c}$, this will be equilibrium threshold. Note that it is possible that $\hat{c}_{AI} < \underline{c}$ if $\alpha V + u - C < \underline{c}$. In this case, no contributors will answer queries in equilibrium. However, this also implies that $\underline{c} > \hat{c}$. Thus, the realised social value is $F(\underline{c})V$, which is again greater than the social value without AI as $\underline{c} > \hat{c}$. ■

The intuition for this result is straightforward. As the AI answers easier queries, then the probability that a random contributor can handle a query falls. Consequently, the expected future payoff for a contributor who does not answer also falls, increasing their incentive to handle a marginally harder set of queries. Thus, with AI, the set of queries handled is $[0, \underline{c}] \cup (\underline{c}, \hat{c}_{AI}] = [0, \hat{c}_{AI}]$ which is larger than the set of queries handled without AI, $[0, \hat{c}]$. This implies that the set of training data expands proportionately to $\hat{c}_{AI} - \hat{c}$. Of course, it is possible that $\underline{c} > \hat{c}_{AI}$. In this case, while the set of training data falls to zero, there are no problems that the AI cannot answer, so social value is higher. Moreover, contributor welfare is also higher as the AI answers more queries, saving on personal contributor costs as $u \leq C$.

The model here assumes that the AI performs without cost, is immediate and provides answers of the same quality as contributors. If AI were costly, took some time and provided answers of lower quality, then these results would continue to hold so long as these differences were not too large. This is because users post queries and will only be satisfied with the AI result if it is of high quality relative to any cost they might incur in using the AI (including paying for the costs of the AI itself). In other words, the user's choice to use an AI determines whether any strict inefficiency associated with AI use translates into actual AI use.

4 Using Pay to Stimulate User-Contributions

The above analysis has shown that while the availability of AI products trained on user contributions may obviate the demand for those user contributions, this substitution does not impact the range of training data generated because it is focused on data that can be more easily generated. When this occurs, user contributions become more concentrated on providing more costly training data and, thus, expand the overall quantum of available training data. Over time, this dynamic, which is powered by selection driven by user intent, will not lead to a dearth of training data; indeed, it will be the opposite.

That said, the analysis here takes place at the level of individual queries. If a platform, such as Stack Overflow, which is advertising-funded, has a significant enough reduction in revenue, it may fail to cover its fixed costs and become unavailable. In this case, some funding from AI providers to support a user contribution platform may be required to sustain the flow of training data. Moreover, training data may be focussed on particular areas, and the improvement of AI products on that dimension may reinforce that focus. For instance, computer languages that have a greater user share may experience stronger network effects as a result of a virtuous circle, while those with a lower user share may experience a strong, vicious circle. This may create "algorithmic monocultures" (Kleinberg and Raghavan, 2021) that may be less robust to changes and harm overall decision-making.

One thing the use of user contributions to train AI products might change is u, which is the intrinsic reward contributors obtain. The model above treats this as an exogenous parameter, but it could be related to design choices that the contribution platform implements (Barbosu and Gans, 2022). For instance, it may be determined by reputational tokens that users bestow on contributors. If an AI answers a significant number of queries, then the allocation of those tokens will change. This could increase the reward per answer or decrease it if fewer users awarding tokens. If u were to decline significantly, either because of design issues or perhaps user repugnance involving in 'gifting' their contributions as training data to commercial AI providers, then there is a sense in which AI availability may cause a reduction in the generation of training data and end up harming overall quality and social welfare.

If this reduction in voluntary contributions to generate training data were to arise, a commercial AI provider may, instead, offer monetary payments as a substitute incentive mechanism. To see how this might work, suppose that an AI provider is able to pay for accurate answers to user queries. We will assume that this involves ideal conditions whereby the provider can perfectly evaluate whether an answer is accurate ex post (e.g., by running the suggested code). Suppose that the pay is w per answer. Suppose also that $u = 0$ in this

environment.

As the answers are still a public good, they are valuable if provided directly or used to train an AI that can provide them. Thus, as before, from a contributor perspective, an answer provided is valued at αV regardless of whether they, someone else or an AI provides it. Given this, the contribution threshold that determines whether a selected contributor will answer a query becomes:

$$\alpha V + w - (C + \hat{c}_{AI}(w)) = \delta \frac{F(\hat{c}_{AI}(w)) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta} \alpha V$$

This is the same as before, except that u has been substituted for w.

Suppose that an individual answer used for training is valued by an AI provider at R in terms of increased revenue. First, note that if $\underline{c} > \alpha V - C$, then increasing w to $C + \underline{c} - \alpha V$ is necessary for any contributor to answer the query and provide data for training. This will only be worthwhile if $R \geq C + \underline{c} - \alpha V$. However, while necessary, these conditions are not sufficient for wages to generate contributions.

Second, note that if $\underline{c} \leq \alpha V - C$, holding $\hat{c}_{AI}$ fixed, offering a higher w can raise the probability a selected contributor provides an answer. This accelerates the provision of answers. This implies that ex ante, for a given query, the AI provider's expected discounted profit is given by:

$$\frac{F(\hat{c}_{AI}(w)) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta} (R - w)$$

Note, however, that for a given w, $\hat{c}_{AI}(w)$ is (from the contribution equation) given by:

$$\hat{c}_{AI}(w) = w - C + \left(1 - \delta \tfrac{F(\hat{c}_{AI}(w)) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta}\right) \alpha V$$

Solving for w and substituting in, we obtain the AI provider's profits as:

$$\tfrac{F(\hat{c}_{AI}(w)) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta} \left(R - \left(\hat{c}_{AI}(w) + C - \left(1 - \delta \tfrac{F(\hat{c}_{AI}(w)) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta}\right) \alpha V\right)\right)$$

For convenience, assume that the marginal acceleration as w rises is given by:

$$\begin{aligned} A(w) &\equiv \frac{\partial \left(\frac{F(\hat{c}_{AI}(w)) - F(\underline{c})}{1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta}\right)}{\partial \hat{c}_{AI}} \frac{d\hat{c}_{AI}}{dw} \\ &= \frac{(1-\delta) f(\hat{c}_{AI})}{(1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta)^2} \frac{1}{1 + \frac{(1-\delta) f(\hat{c}_{AI})}{(1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta)^2}} \\ &= \frac{(1-\delta) f(\hat{c}_{AI})}{(1-\delta) f(\hat{c}_{AI}) + (1 - (1 - (F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta)^2} \end{aligned}$$

Then, the marginal effect on profits from increasing w is:

$$A(w)(R - \hat{c}_{AI} - C + \alpha V) - \frac{F(\hat{c}_{AI}(w)) - F(\underline{c})}{1-(1-(F(\hat{c}_{AI}(w)) - F(\underline{c})))\delta} \frac{d\hat{c}_{AI}}{dw} - \delta A(w)^2 \alpha V$$

The last two terms are negative. Thus, while a higher wage does accelerate performance, it does so for other contributors as well, which has the inframarginal effect of decelerating performance. Without further information, it is not clear whether an AI provider would choose to pay for training data generation.

To consider this further, suppose that the AI provider starts from a position where $w = 0$. In this case, the goal is to create conditions under which $\hat{c}_{AI}(w) > \underline{c}$. Note that at $\hat{c}_{AI} = \underline{c}$, $A(0) = \frac{f(\underline{c})}{1-\delta+f(\underline{c})}$. Therefore, the marginal effect of w on profits at $\underline{c}$ is:

$$\frac{f(\underline{c})}{1-\delta+f(\underline{c})} \left(R - \underline{c} - C + (1 - \delta \frac{f(\underline{c})}{1-\delta+f(\underline{c})}) \alpha V \right)$$

This is positive if and only if:

$$\underline{c} < R - C + \frac{(1-\delta)(1+f(\underline{c}))}{1-\delta+f(\underline{c})} \alpha V$$

This is compared to the condition whereby contribution is efficient if $\underline{c} < R - C + \alpha V$. Therefore, there exist some efficient contributions that the provider will not end up paying for (that is, where the marginal effect of w on profits is negative).

5 Conclusion

The generation of training data is a critical component in developing and improving AI prediction algorithms. This paper has examined one of the economic considerations that arise in this process. Using user-contributed data for AI training does not necessarily reduce the quantity or quality of such data over time. Although AI may substitute for some user contributions, it is likely to focus on easier queries, leading to a greater concentration of user efforts on more complex and valuable contributions. In addition, the scope for explicit pay to alleviate concerns that may arise in terms of the overall ongoing scale of contributions is shown, by the same mechanism, to be muted in scope.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Barbosu, S. and Gans, J. S. (2022). Storm crowds: Evidence from zooniverse on crowd contribution design. *Research Policy*, 51(1):104414.

Borwankar, S. and Khern-am nuai, W. (2023). Unraveling the impact: An empirical investigation of chatgpt's exclusion from stack overflow. *Available at SSRN 4481959*.

del Rio-Chanona, M., Laurentsyeva, N., and Wachs, J. (2023). Are large language models a threat to digital public goods? evidence from activity on stack overflow. *arXiv preprint arXiv:2307.07367*.

Easley, D. and Ghosh, A. (2016). Incentives, gamification, and game theory: an economic approach to badge design. *ACM Transactions on Economics and Computation (TEAC)*, 4(3):1–26.

El-Komboz, L. A. and Goldbeck, M. (2024). Career concerns as public good-the role of signaling for open source software development. Technical report, ifo Institute-Leibniz Institute for Economic Research at the University of ....

Engers, M. and Gans, J. S. (1998). Why referees are not paid (enough). *American Economic Review*, 88(5):1341–1349.

Gallea, Q. (2023). From mundane to meaningful: Ai's influence on work dynamics–evidence from chatgpt and stack overflow. *arXiv preprint arXiv:2308.11302*.

Ghosh, A. (2013). Game theory and incentives in human computation systems. In *Handbook of Human Computation*, pages 725–742. Springer.

Ghosh, A. and Hummel, P. (2011). A game-theoretic analysis of rank-order mechanisms for user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 189–198.

Ghosh, A. and McAfee, P. (2011). Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, pages 137–146.

Kleinberg, J. and Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

Lerner, J. and Tirole, J. (2002). Some simple economics of open source. *Kournal of Industrial Economics*, 50(2):197–234.

Somers, J. (2023). How will a.i. learn next? *The New Yorker*.

Xu, B., Nguyen, T.-D., Le-Cong, T., Hoang, T., Liu, J., Kim, K., Gong, C., Niu, C., Wang, C., Le, B., et al. (2023). Are we ready to embrace generative ai for software q&a? In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1713–1717. IEEE.

Xu, L., Nian, T., and Cabral, L. (2020). What makes geeks tick? a study of stack overflow careers. *Management Science*, 66(2):587–604.