# EXHIBIT 40

## An Economic Model for Compensating Data Producers

One of the controversies over Copilot, related to but distinct from concerns about the model spitting out copyrighted data, is a feeling that large corporations (MSFT and OpenAI) are exploiting the open source work of developers and using it extract concentrated profits. The same could be said about GPT-3 being trained partially on fanfiction books, or for that matter about models being trained on human feedback, for which workers are paid near-minimum wage. I think this is a real and important concern, especially when considered as part of an overall trend for AI progress to concentrate wealth, increase inequality, and in the long run automate labor. Zooming very far out, what is happening macroeconomically is that distributed human labor is being used to train AI models by large, centralized actors, who concentrate the resulting profits while in the long run making the human labor obsolete. It seems like the default trajectory is for one of the following things to happen:

- The current trend continues, and AI model training becomes an increasingly extractive concentrator of wealth, in more and more industries like writing, music, video-making, learning of various human interactive skills, etc. Content creators grumble and AI companies become more and more unpopular, but the trend of increasingly powerful models, increasing automation, and increasing concentration of wealth continues.
- Content creators (programmers, artists, etc) get mad and act to prevent their work from being used for AI model training. New open source legal licenses are created which resemble current open source licenses with the one modification that they explicitly ban the use of their data for AI model training. The extractive economic process is stopped, but at the cost of drastically slowing down AI progress (including both safety and non-safety work).
- The government gets mad and cracks down on AI companies, probably not over this issue in particular, but over a number of issues including this.

**Commented [1]:** IANAL, but I don't think this would work. Training models on data doesn't rely on the license, it relies on jurisprudence that training models is fair use. In theory, I believe that we could train our models on proprietary code if we had access to it (although if our model regurgitated the proprietary code, we'd have much worse liability).
[Author:Christopher Olah],[Creator:1297307460],[Date:7/12/2021 4:04:00 PM]

**Commented [2]:** Maybe you could try to make it a TOS for accessing the code? "By downloading this code, you agree to not train neural networks on it..." But my sense is that website TOS like that generally haven't turned out to be enforceable.
[Author:Christopher Olah],[Creator:1297307460],[Date:7/12/2021 4:05:00 PM]

**Commented [3]:** I think the only way to actually change things would be legal decisions limiting when training models is fair use or legislation.
[Author:Christopher Olah],[Creator:1297307460],[Date:7/12/2021 4:06:00 PM]

**Commented [4]:** Again, I'm not a lawyer. Would be interesting to have a professional view.
[Author:Christopher Olah],[Creator:1297307460],[Date:7/12/2021 4:06:00 PM]

None of these outcomes seem particularly good for most people. **An alternative that may work better for everyone, is to compensate data/content producers for their labor, not at a flat rate, but with a fraction of the profits from the model produced.** Essentially, we can think of the trained AI model as a corporate entity, and the data/content creators own equity in it by virtue of their contribution to training it. Initially, companies (such as Anthropic, for which this would be a great fit for our public benefit orientation) could voluntarily grant such equity as a good deed, but over time it would evolve into a norm and perhaps eventually a legal right. Distribution would of course be a problem, but could be handled in the short run through the platforms on which content is created (e.g. GitHub, fanfiction websites) and in the long run though some dedicated, Patreon-like platform.

The other major question is how to value the contribution of data creators to AI models -- what fraction of the model should they own, collectively and individually? Here is where the technical work of people on our team may have surprising relevance. If we want to know how important data vs compute vs algorithms are as a factor of the performance of any given AI model, this is precisely the question which scaling laws answer. Specifically:

- We can calculate the marginal impact of data vs compute on the performance of the model, using the standard scaling laws, to get a “marginal utility of data in loss units”, which would be the nominal price in a marketplace.
- Probably this undervalues data relative to what we’d want in some common-sense framework -- if data is very abundant the marginal value could be basically zero, and we care about something more like the average value (rather than marginal), although this is hard to define since data, compute, and algorithms are all necessary ingredients. We could declare data to be worth some flat amount -- like 20% -- or we could make some attempt to calculate the average.
- We’d also want to convert loss units to profit units -- essentially, how much more profitable is a model with loss X compared to a model with loss e.g. 1.1X? This can likely be answered empirically, or we could make some curve where models with loss above a certain value are worth zero, and we just extrapolate from there.
- This all relates to the bulk value of data. What about the individual value? The easiest way is just to compensate by number of words/tokens, in other words to assume all data has the same value and distribute that value in proportion to the raw amount of data contributed. You could also try more sophisticated ideas like attempting to measure the value (though the size of the gradient update??) of individual pieces of data using the model itself. Though this could get weird if some data ends up being worth much more than other data for inscrutable algorithmic reasons.
- To do a basic check on the amounts, we can easily imagine a very powerful coding model in the future producing $1B in profit every year when deployed by a large company. If this model is trained on the work of 1 million programmers, and data is valued at 20% of the equity in the model, that’s $200/year for the average programmer, per model deployed. It’s unclear how many such models will be deployed, but it could easily be 10, 20, or more, which could take us up to the level of ~$2,000-$4,000 a year -- not trivial by any means, and probably well worth doing. Programmers are relatively wealthy, but for other forms of content (e.g. music or art), these amounts could be even more meaningful.

All of this is a bit of a crazy idea, and has the potential to end up looking weird if it compensates people in an unfair or inscrutable way. If done ham-handedly it could also create privacy issues (similar to what the “Worldcoin” idea is experiencing), or create the impression that people are being crudely bribed. But if done thoughtfully, it might be a solution everyone likes.

In the long run, we can imagine a world in which AI agents will effectively be amalgams or representatives of the values and skills of millions of people, who in turn own the economic output of these agents, no matter how distantly removed they become from the original human behaviors that produced them -- a sort of “guaranteed basic capital” for the machine age.

**Commented [5]:** The principled way to do this is to use influence functions. This would allow you to ask "how much did data X influence prediction Y on the margin?" So you could say "well, this was a small amount of data, but influenced really important answers." But computing it would likely be prohibitively expensive. [Author:Christopher Olah],[Creator:1297307460],[Date:7/12/2021 4:13:00 PM]

**Commented [6]:** (Doing full dataset to many predictions computation would cost many fold the cost of training the model!) [Author:Christopher Olah],[Creator:1297307460],[Date:7/12/2021 4:13:00 PM]

**Commented [7]:** A cheaper way: if we do a model that relies on retrieval, you could price data partly by the economic value of cases where it is retrieved, and possibly do attribution on top of that to pin down influence even more. This would be extremely viable to compute. [Author:Christopher Olah],[Creator:1297307460],[Date:7/12/2021 4:14:00 PM]

**Commented [8]:** Up front: I think this is a fantastic discussion to have, and this is a great pitch. I think it gives a plausible mechanism for us to come up at a price at which we'd *buy* data, but to make a market it needs to be paired with a mechanism for figuring out what price people'd *sell* their data at.

Without that, you're basically declaring the value of people's labour by fiat. I think that's compatible wi ... [1]

**Commented [9]:** That all makes sense. I think we're not going for a true market equilibrium (the true equilibrium price could be close to zero if the marginal value of the data is close to zero -- economics does not make reward proportional to effort, nor does it necessarily produce "fair" outcomes), but a sort of quasi-market equilibrium designed to artfully counter the concentration of wealth. ... [2]

**Commented [10]:** On the profits getting driven to zero -- it gets a bit complicated, but if AI offspring outcompete human data generators, isn't it fine as long as those AI offspring are themselves owned by the humans who originally generated *them*? As long as the thread goes back to a human data generator somewhere, then whatever profits are produced are ultimately owned by humans, at however far a ... [3]

**Commented [11]:** Think it'd end with providing everything to people at the cost of the physical inputs - compute infrastructure and energy. Farming or forestry might be a good comparison here, where there are lots of providers of substitutable goods and yeah, no great concentration of wealth.

Some quick thoughts about the sell-price mechan ... [4]

**Page 2: [1] Commented [8] Andy Jones 7/12/2021 2:40:00 PM**

Up front: I think this is a fantastic discussion to have, and this is a great pitch. I think it gives a plausible mechanism for us to come up at a price at which we'd *buy* data, but to make a market it needs to be paired with a mechanism for figuring out what price people'd *sell* their data at.

Without that, you're basically declaring the value of people's labour by fiat. I think that's compatible with many ideologies - both past and future - but possibly not with ours.

The other more market-oriented worry is that I've got an intuition that multiple AI companies competing on price'd drive profits to zero, along with the payments to the data-generators. Who are then outcompeted by their AI offspring.

A mechanism for figuring out the price at which people'd sell could be as simple as a government-declared minimum price, or as complex as... well, it could get pretty complex. [Author:Andy Jones],[Creator:1297307460],[Date:7/12/2021 2:40:00 PM]

**Page 2: [2] Commented [9] Dario Amodei 7/12/2021 3:04:00 PM**

That all makes sense. I think we're not going for a true market equilibrium (the true equilibrium price could be close to zero if the marginal value of the data is close to zero -- economics does not make reward proportional to effort, nor does it necessarily produce "fair" outcomes), but a sort of quasi-market equilibrium designed to artfully counter the concentration of wealth.

On the buy vs sell, I agree that's a good point. Having only one side of the market is like forcefully buying the data by eminent domain. I suppose the price for selling comes from the implicit threat content providers have to disallow training on their data -- not sure how we would calculate that, but seems worth doing something like it.

The whole equilibrium might in fact just work itself out, given enough time, by this very mechanism, but not without a bunch of hatred, threats, and mutual acrimony. Or it might not work itself out due to collective action problems or blunt government action. The hope here would be to shortcut all that and just propose the equilibrium solution at the start. [Author:Dario Amodei],[Creator:1297307460],[Date:7/12/2021 3:04:00 PM]

**Page 2: [3] Commented [10] Dario Amodei 7/12/2021 3:07:00 PM**

On the profits getting driven to zero -- it gets a bit complicated, but if AI offspring outcompete human data generators, isn't it fine as long as those AI offspring are themselves owned by the humans who originally generated *them*? As long as the thread goes back to a human data generator somewhere, then whatever profits are produced are ultimately owned by humans, at however far a remove. If all the profits in the economy go to zero -- then I'm not sure what happens. Naively it seems like you have all these AI's running around providing everything to people for free, and no accompanying concentration of wealth? [Author:Dario Amodei],[Creator:1297307460],[Date:7/12/2021 3:07:00 PM]

**Page 2: [4] Commented [11] Andy Jones 7/12/2021 3:41:00 PM**

Think it'd end with providing everything to people at the cost of the physical inputs - compute infrastructure and energy. Farming or forestry might be a good comparison here, where there are lots of providers of substitutable goods and yeah, no great concentration of wealth.

Some quick thoughts about the sell-price mechanism:
 * Brave (the browser) is a recent example of trying to build market mechanisms around a recently-valuable resource that everyone has a supply of.
 * This also feels like it has shades of all the old hard-to-price resource problems like 'rights to drill for oil on someone's land' or 'water rights' or 'blood donation'. The solutions we came up with there are pretty bad and a long, long way from the nigh-on-frictionless solution we need here. Honestly, this might be an academic economist's field day. Do we have anyone economy-y at FHI we're buds with?
 * I have some ill-thought out ideas about AI factors that represent you in negotiations with other AI factors. Because when you've got a hammer, [Author:Andy Jones],[Creator:1297307460],[Date:7/12/2021 3:41:00 PM]