Should We Treat Data as Labor?

Moving Beyond "Free"

By IMANOL ARRIETA IBARRA, LEONARD GOFF, DIEGO JIMÉNEZ HERNÁNDEZ, JARON LANIER, AND E. GLEN WEYL *

In the previous paper in this session and in a forthcoming book (Posner and Weyl, 2018), one of us argues that by creating or strengthening absent markets, we can simultaneously address the inequality, stagnation and sociopolitical conflict afflicting developed countries. He calls such cases "radical markets" because of their transformative emancipatory potential. A promising example was suggested years earlier by another of us, who wrote a book (Lanier, 2013) highlighting the social problems with the culture of "free" online, in which users are neither paid for their data contributions to digital services nor pay directly for the value they receive from these services. While free data for free services is a barter, he argued that the lack of targeting of incentives undermines market principles of evaluation, skews distribution of financial returns from the data economy and stops users from developing themselves into "first-class digital citizens". In this paper we explore whether and how treating the market for data like a labor market could serve as a radical market that is practical in the near term.

I. The High Cost of Free Data

The digital economy is perhaps the leading source of innovation today, delivers massive surplus to users (Brynjolffson et al., 2017) and is "free" (at point of use) to users. Despite these benefits, popular anxiety and backlash is rising.

The most common concern is employment and income distribution. Many fear that artificial intelligence (AI) systems will replace human workers. Economists rightly respond that greater technological disruptions in the past, while causing shifts in employment, have largely left labor's share of income constant or even growing (Autor, 2015). Yet recent secular declines in labor's share (Karabarbounis and Neiman, 2014) belie its universal stability.

Furthermore, the employment numbers of leading technology companies give little cause for optimism. The market capitalization and value-added of firms like Facebook, Google and Microsoft are similar to or greater than a firm like Walmart, yet they employ 1-2 orders of magnitude fewer workers and our primitive attempts to estimate the labor income shares of these companies from publicly available statistics suggest they are a small fraction of the traditional average 60-70%. The "future" such firms represent would validate Piketty (2013)'s foreboding of high capital shares.

Simultaneously, the lack of payment to users for data may drag on the contributions of AI to productivity growth. Despite the widespread hype about AI, its contributions to productivity seem to have been limited thus far (Gordon, 2016; Nadella, 2017). A potential explanation relates to the role of data. The first generation of AI systems largely failed to achieve their goals because they relied too heavily on hard-coding by engineers. The new generation of AI uses statistical methods called "machine learning" (ML), which adapt to patterns in examples of humans performing similar tasks ("big data").

Yet the free data model has made productivity-related data much less accessible than consumption-oriented data. Workers who expect to be compensated are the primary

* Arrieta: Department of Management Science and Engineering, School of Engineering, Stanford University, Huang Engineering Center, 475 Via Ortega Avenue, Stanford, CA 94305 (imanol@stanford.edu). Goff: Department of Economics,Columbia University, 1022 International Affairs Building, 420 West 118th Street, New York, NY 10027 (ltg2111@columbia.edu). Jiménez: Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305 (diego.jimenez@stanford.edu). Lanier: Office of the Chief Technology Officer, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 (jalani@microsoft.com) Weyl: Microsoft Research, One Memorial Drive, Cambridge, MA 02142 and Yale University Department of Economics and Law School (glenweyl@microsoft.com). We are grateful to many colleagues for comments, but especially to Microsoft business leaders Satya Nadella and Kevin Scott for their encouragement. All errors are our own.

performers of productivity-related tasks and these often occur within firms unwilling to surrender their proprietary internal data to AI companies for free. More broadly, many AI systems depend on active participation by humans to generate relevant data. This ranges from users granting permission to access data naturally created in the course of consumption experiences, through users that go out of their way to provide examples of translations or feedback on translations generated by AIs as they use these systems, to the sort of active labeling and analysis tasks currently supplied in digital labor markets such as Amazon's Mechanical Turk or Mighty AI (Gray and Suri, 2017) and even to the creative content displayed on blogs and video sharing sites.

However, these systems seem inefficient as they generally do not reward those with the greatest expertise and context (usually those producing the data that others currently label in the first place), either reassigning task to those with little context or coaxing those with context to provide feedback for free as part of accessing online services (as in the case of DuoLingo or reCAPTCHA). They appear to be workarounds to avoid directly paying those best able to supply high-quality data rather than efficient procurement practices. A purely free data economy acts as a drag on productivity growth that continues to lag worldwide (Byrne et al., 2016) despite bold hopes for AI's potential.

Finally, recent anxiety about employment and the digital economy goes beyond the purely economic. On the one hand, increasing numbers of workers, especially away from cosmopolitan and high-tech cities, are disillusioned with and disenfranchised by technological and economic progress. Many believe these feelings helped stimulate populist movements of the left and right throughout the developed world.

Simultaneously young people spend increasing time on and have developed increasing expertise in digital interactions such as social media and video games (Perrin, 2015; Aguiar et al., 2017). Because such activities are overwhelmingly framed as consumption rather than production, these growing online lives are widely seen as running contrary to or undermining the dignity provided by work. Many of these young people seem to have become involved with antisocial activities (such as cyberbullying and hate speech) or to have declining self-esteem. Thinkers promoting the idea of a "universal basic income (UBI)" have even suggested dignity based on work is becoming outdated and that as AI replaces humans leisure may be a growing source of identity (Parijs and Vanderborght, 2017). Whatever the promise of this idea, for the medium term treating online experiences as purely consumption holds risks for the social and political fabric of developed countries.

II. Capital or Labor?

We contend that the key aspect of the current political economy of data that causes these problems is treating data as capital rather than as labor. While it might seem that assets either are one or the other, and that treatment is irrelevant, transitions in the social attitude towards assets across these categories have played important roles in history. Slavery and to a lesser extent feudalism treated (largely agricultural) work as a possession of a master or lord, while liberal and labor reform worked to give recognition and its marginal economic product to labor. To understand what we are trying to accomplish, it is useful to contrast several attitudes towards data at present under the "Data as Capital (DaC)" paradigm to those appropriate in a world where we see data as labor (DaL); we summarize these in Table 1.

DaC treats data as natural exhaust from consumption to be collected by firms, while DaL treats them as user possessions that should primarily benefit their owners. DaC channels payoffs from data to AI companies and platforms to encourage entrepreneurship and innovation, while DaL channels them to individual users to encourage increased quality and quantity of data. DaC prepares for AI to displace workers either by supporting UBI or reserving spheres of work where AI will fail for humans, while DaL sees ML as just another production technology enhancing labor productivity and creating a new class of "data jobs". DaC encourages workers to find dignity in leisure or in human interactions outside the digital economy, while DaL views data work as a new source of "digital dignity". DaC sees the online social contract as free services in exchange for prevalent surveillance, while DaL sees the need for large-scale institutions to check the ability of data platforms

Issue	Data as Capital	Data as Labor
Ownership	Corporate	Individual
Incentives	Entrepreneurship	"Ordinary" contributions
Future of work	Universal Basic Income	Data work
Source of self-esteem	Beyond work	Digital dignity
Social contract	Free services for free data	Countervailing power to create data labor market

TABLE 1—LEADING CHARACTERISTICS OF THE "DATA AS CAPITAL" VERSUS "DATA AS LABOR" PERSPECTIVES.

to exploit monopsony power over data providers and ensure a fair and vibrant market for data labor.

Describing DaL versus DaC as a binary is obviously too simplistic and extreme. Production function for data and the AI systems built on top of it are certainly more continuous: data, capital (e.g. computational power), skilled labor (e.g. programmers), entrepreneurial talent and "land" (e.g. rents on network effects) all matter and these different inputs can likely be substituted reasonably smoothly. The socially optimal shares of each factor depends on as-yet-unmeasured details of production functions and data themselves are not purely created by users: they requires firms to track, record and organize user behavior.

Yet we doubt the optimal (viz.competitive) share of user data contributions is a negligible fraction of the total value of the digital economy. While the marginal value of data in estimating any finite dimensional quantity eventually steeply declines, the power of the latest generation of ML has been its ability to tackle increasingly sophisticated tasks as the quality and quantity of data improve. Many of these more sophisticated tasks are impossible to even get started on without ample data, as the neural networks and other learning algorithms required cannot learn the right representations of complex phenomena without many training examples. This suggests that the returns to data may decline only gradually or there may even be increasing returns to data if more sophisticated tasks are disproportionately more valuable. This is consistent with the empirically-observed dominance of the data economy by a few large firms.

Luckily, the production function for AI may be easier to measure than other production functions because the relevant ML algorithms and their performance at different times and for different data sets are usually well-documented, at least internally to companies. Combining these with advances in ML that allow estimation of the marginal effect of new data on predictions (Koh and Liang, 2017) suggests a promising avenue for valuing data (and one we are pursuing at Microsoft), though there are many conceptual and computational challenges still to be overcome.

Whatever the precise balance, the only "third way" out of the DaL-DaC spectrum we see is the failure of AI: if AI proves to be relatively unproductive or irrelevant, neither DaL nor DaC will much matter. But if AI lives up to even a part of its hype, failure to move towards DaL will leave us trapped in the problems we highlight with DaC.

III. How Did We Get Here?

If treating data purely as capital is economically and socially irrational, how have we ended up in the present equilibrium? As in the nineteenth century labor struggles, the usual culprits are a combination of prejudice (viz. the weight of precedent created by historical accidents) and privilege (viz. entrenched interests that derive rents from the inefficient equilibrium). In the present setting, user expectations of "lightweight" online experiences has conspired with the monopsony power of the technology giants (what one of us has called "siren servers") to maintain the status quo.

The internet economy largely began with a venture-capital fueled bubble that chased usage with little sense for a business model. The social movement for "free software" collided with a counter-cultural streak in Silicon Valley that declared information wants to be free and built users expectations of digital services being offered freely. Searching for a way to monetize this activity, Google and then Facebook turned to advertising targeted using user data. This accustomed users to surrendering data in exchange for free services (Carrascal et al., 2013), expectations that have persisted as the value of such data to broader AI services has risen. Few users

are even aware of the productive value of their data or the role they play in enabling ML.

Yet historical accidents have not only entrenched expectations and norms, they also have created powerful interests in maintaining the *status quo*. The largest siren servers, especially Facebook and Google, but also Microsoft and others, benefit from the free or extremely cheap availability to them of data. While the total value created by data might be much larger in a DaL world, users aware of the value of their data would likely demand compensation in a range of settings, dramatically reducing the share of value that could be captured by the siren servers as profits. This is just an extreme version of the standard logic of monopsony: while a usual monopsonist just depresses wages, the historical background we explain above has made it attractive for siren servers to maintain a DaC equilibrium where users are not even aware of the value their data daily create for siren servers.

Recent evidence suggests significant monopsony power in online task labor markets. Dube et al. (2018) use randomly varied wages on Amazon Mechanical Turk to find elasticities of the labor supply curve facing a task-poster that are well below unity. These small task-posters almost certainly have more elastic residual labor supply than does a siren server, suggesting extreme monopsony power in the latter case: a question we have been investigating in on-going work with Microsoft data. In on-going work using a large Microsoft program that pays users in loyalty points for Bing searches, we estimate even smaller elasticities in the number of searches performed among active users of the program. This reinforces the idea that monopsony may be an important force blocking the potential productivity gains from DaL

IV. Sources of Countervailing Power

The inefficient exploitation of labor by concentrated capital was a constant theme of political economy before the Cold War. Galbraith (1952) summarized various solutions to this problem as forms of "countervailing power" by large scale social institutions.

In the data economy, the first and most natural balancing factor is competition. While Facebook and Google rely heavily on DaC, other leading technology companies (e.g. Amazon and Apple) mostly follow different business models and a productivity-oriented company like Microsoft might even benefit from users perceiving themselves more as producers online. These other companies also lag Facebook and Google in the data race to train ML systems. Returning more of the gains to data laborers might help them compete in creating AI systems. Smaller companies or start-ups could also make a difference, and many (e.g. Meeco) have been formed around DaL-related ideas. Yet we doubt, given the economies of scale related to data in producing AI systems, that a smaller player could succeed without a significant partnership with one of the largest technology companies.

Second, data laborers could organize a "data labor union" that would collectively bargain with siren servers. While no individual user has much bargaining power, a union that filters platform access to user data could credibly call a powerful strike. Such a union could be an access gateway, making a strike easy to enforce and on a social network, where users would be pressured by friends not to break a strike, this might be particularly effective. A union could also be useful in certifying data quality and guiding users to develop their earning potential.

Finally, governments can play an important role in helping facilitate DaL both on the positive and negative side. On the positive side, new regulatory frameworks such as the European General Data Protection Regulations are increasingly shifting ownership rights in data to the users that generate them. Data collectors increasingly must allow users to understand, withdraw and transfer their data across competitors. On the other hand, existing labor laws fit poorly with a world where much data labor may be done in the course of consumption experiences rather than as a dedicated activity. Adapting labor laws to defend workers against monopsony while allowing the flexibility data work will require a combination of economic and technical sophistication that we hope labor economists can increasingly provide to support policy-makers.

V. A Radical Data Market

Ultimately, we believe all three of these factors must coordinate for DaL to succeed, just

as in historical labor movements. Whatever the mix, however, building a market for data labor offers economists an exciting chance to design a market on a much broader scale than most work on market design in the past (Roth, 2015). For example, we are currently working to use regularized measures of the marginal value of data points to design and make transparent efficient payments for data workers. With studies projecting that AI might automate as many as 50% of jobs in the coming decades (Frey and Osborne, 2017), data labor has the potential to constitute a significant fraction of national income. At the same time, economists, in their roles as advisors to governments and technology companies, are likely to play a central role in defining the texture of these markets. A radical market in data labor offers a near-term opportunity for economists, in collaboration with the other social and computer scientists they regularly work with in the technology industry, to bring years of research in labor economics and market design to bear on a central social problem of our times.

REFERENCES

Aguiar, Mark, Mark Bils, Kerwin Kofi Charles, and Erik Hurst, "Leisure Luxuries and the Labor Supply of Young Men," 2017. http://www.nber.org/papers/w23552.

Autor, David H., "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *Journal of Economic Perspectives*, 2015, *29* (3), 3–30.

Brynjolffson, Erik, Felix Eggers, and Avinash Gannamaneni, "Using Massive Online Choice Experiments to Measure Changes in Well-being," 2017. Latest version available from authors.

Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf, "Does the United States have a Productivity Slowdown or a Measurement Problem?," *Brookings Papers on Economic Activity*, 2016, (Spring), 109–182.

Carrascal, Juan Pablo, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira, "Your Browsing Behavior for a Big Mac: Economics of Personal Information Online," in "Proceedings of the 22Nd International Conference on World Wide Web" WWW '13 ACM New York, NY, USA 2013, pp. 189–200.

Dube, Aindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri, "Monopsony in Online Labor Markets," 2018. This paper is under preparation. Contact Suresh Naidu at suresh.naidu@gmail.com for a copy.

Frey, Carl Benedikt and Michael A. Osborne, "The Future of Employment: How Susceptible are Jobs to Computerisation?," *Technological Forecasting and Social Change*, 2017, *114*, 254–280.

Galbraith, John Kenneth, *American Capitalism*, New York: Houghton Mifflin, 1952.

Gordon, Robert J., *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*, Princ, 2016.

Gray, Mary L. and Siddharth Suri, "The Humans Working Behind the AI Curtain," *Harvard Business Review*, January 9 2017.

Karabarbounis, Loukas and Brent Neiman, "The Global Decline of the Labor Share," *Quarterly Journal of Economics*, 2014, *129* (1), 61–103.

Koh, Pang Weh and Percy Liang, "Understanding Black-Box Predictions Via Influence Functions," in "Proceedings of Machine Learning Research," Vol. 70 2017, pp. 1885–1894.

Lanier, Jaron, *Who Owns the Future?*, New York: Simon & Schuster, 2013.

Nadella, Satya, *Hit Refresh: The Quest to Rediscover Microsoft's Soul and Imagine a Better Future for Everyone*, New York: Harper Business, 2017.

Parijs, Philippe Van and Yannick Vanderborght, *Basic Income: A Radical Proposal for a Free Society and a Sane Economy*, Cambridge, MA: Harvard University Press, 2017.

Perrin, Andrew, "Social Media Usage: 2005-2015," Technical Report, Pew Research Center 2015.

Piketty, Thomas, *Le Capital au XXI^e Siècle*, Paris: Éditions du Seuil, 2013.

Posner, Eric A. and E. Glen Weyl, *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*, Princeton, NJ: Princeton University Press, 2018.

Roth, Alvin E., *Who Gets What – and Why: The New Economics of Matching and Market Design*, New York: Houghton Mifflin Harcourt, 2015.