WHOSE DATA, WHOSE VALUE?

SIMPLE EXERCISES IN DATA AND MODELING EVALUATION WITH IMPLICATIONS FOR TECHNOLOGY LAW AND POLICY

AILEEN NIELSEN*

Scholarship on the phenomena of big data and algorithmically-driven digital environments has largely studied these technological and economic phenomena as monolithic practices, with little interest in the varied quality of contributions by data subjects and data processors. Taking a pragmatic, industry-inspired approach to measuring the quality of contributions, this work finds evidence for a wide range of relative value contributions by data subjects. In some cases, a very small proportion of data from a few data subjects is sufficient to achieve the same performance on a given task as would be achieved with a much larger data set. Likewise, algorithmic models generated by different data processors for the same task and with the same data resources show a wide range in quality of contribution, even in highly performance-incentivized conditions. In short, contrary to the trope of data as the new oil, data subjects, and indeed individual data points within the same data set, are neither equal nor fungible. Moreover, the role of talent and skill in algorithmic development is significant, as with other forms of innovation. Both of these observations have received little, if any, attention in discussions of data governance. In this essay, I present evidence that both data subjects and data controllers exhibit significant variations in the measured value of their contributions to the standard Big Data pipeline. I then establish that such variations are worth considering in technology policy for privacy, competition, and innovation.

The observation of substantial variation among data subjects and data processors could be important in crafting appropriate law for the Big Data economy. Heterogeneity in value contribution is undertheorized in tech law scholarship and implications for privacy law, competition policy, and innovation. The work concludes by highlighting some of these implications and posing an empirical research agenda to fill in information needed to realize policies sensitive to the wide range of talent and skill exhibited by data subjects and data processors alike.

 Visiting Assistant Professor of Law, Harvard Law School. The author gratefully acknowledges financial support from the Program on Economics & Privacy (PEP) at George Mason University. For helpful comments and critiques, the author thanks James Cooper, Erika Douglas, Lilian Edwards, Jeanne Fromer, and Salome Viljoen, as well as conference participants at PEP's Research Roundtable on Regulating Privacy.

INTRODUCTION

2022 gave the world an artificial intelligence (AI) winter in a very new sense of the term.[1] On November 30, 2022, OpenAI launched ChatGPT,[2] a novel and delightfully usable interface powered by the GPT-3.5 series large language model (LLM).[3] The launch of ChatGPT came on the heels of other splashy announcements in the generative AI industry earlier in 2022,

---

[1] *See* Ellen Glover & Brennan Whitfield, *What Is AI Winter?*, BUILT IN (Apr. 21, 2023), https://builtin.com/artificial-intelligence/ai-winter [https://perma.cc/R3DL-KADP]. The first and second AI winters occurred in the mid-1970s and mid-1990s, in each case reflecting a cycle of over-promising causing enthusiasm and funding to wane. *See id.* By contrast, what I am labeling the AI winter of 2022-2023 reflected a period of unprecedentedly high interest in, use of, and (arguably) performance by large language models (LLMs), amazing the general public and alarming some political leaders. *See, e.g.*, Patrick Tucker, *The Pentagon's AI Chief Is 'Scared to Death' of ChatGPT*, DEF. ONE (May 3, 2023), https://www.defenseone.com/technology/2023/05/pentagons-ai-chief-scared-death-chatgpt/385963 [https://perma.cc/4FSH-88HW]. *See generally External Links*, Section of *AI Winter*, WIKIPEDIA (Oct. 30, 2023), https://en.wikipedia.org/wiki/AI_winter#External_links [https://perma.cc/XC6C-94DC].

[2] *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), https://openai.com/blog/chatgpt [https://perma.cc/4XLN-3DJQ].

[3] *Id.*; *see also* Ashley Pilipiszyn, *GPT-3 Powers the Next Generation of Apps*, OPENAI (Mar. 25, 2021), https://openai.com/blog/gpt-3-apps [https://perma.cc/J56M-4BCD].

including OpenAI's launch of DALL-E 2[4] and Stability AI's launch of Stable Diffusion.[5] In the space of a few months, generative AI grew from a potentially interesting, mildly troubling, temporally distant phenomenon to a digital product outcompeting humans for jobs[6] and triggering high stakes litigation by powerful holders of intellectual property rights.[7] As a result of the leap forward in generative AI performance and ease of use, experts anticipate rapid structural shifts in some industries.[8]

Generative AI's sharp turn towards commercial success has brought new urgency to previously slow-burning debates among policymakers and scholars about who owns data and how the economic value created by data-driven AI could or should be distributed among stakeholders. Such stakeholders include parties who, wittingly or otherwise, contributed training data.[9] Data providers—be they data controllers or data subjects—contribute value to generative AI models, often non-consensually.[10] This work broadly

---

4 *DALL-E Now Available Without Waitlist*, OPENAI (Sept. 28, 2022), https://openai.com/blog/dall-e-now-available-without-waitlist [https://perma.cc/2956-B46F].

5 *Stable Diffusion Launch Announcement*, STABILITY.AI (Aug. 10, 2023), https://stability.ai/blog/stable-diffusion-announcement [https://perma.cc/5XXP-TVGA].

6 *See* Martin K.N. Siele, *AI Is Taking the Jobs of Kenyans Who Write Essays for U.S. College Students*, REST OF WORLD (Apr. 21, 2023), https://restofworld.org/2023/chatgpt-taking-kenya-ghostwriters-jobs [https://perma.cc/G6NP-2RVQ] (describing how ChatGPT is reducing earnings of freelancers in the contract cheating industry); *see also* Beatrice Nolan, *Employee Says ChatGPT Carries out 80% of His Work Duties, Which Allowed Him to Take on a 2nd Job, Report Says*, BUS. INSIDER (Apr. 27, 2023), https://www.businessinsider.com/chatgpt-second-job-overworking-overemployment-2023-4 [https://perma.cc/B4YD-CT8L].

7 *See* James Vincent, *Getty Images Is Suing the Creators of AI Art Tool Stable Diffusion for Scraping Its Content*, VERGE (Jan. 17, 2023), https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit [https://perma.cc/RVT8-PZ6H].

8 *See* Beatrice Nolan, *AI Systems Like ChatGPT Could Impact 300 Million Full-Time Jobs Worldwide, with Administrative and Legal Roles Some of the Most at Risk, Goldman Sachs Report Says*, BUS. INSIDER (Mar. 28, 2023), https://www.businessinsider.com/generative-ai-chatpgt-300-million-full-time-jobs-goldman-sachs-2023-3 [https://perma.cc/5UED-EEWJ].

9 *See generally Ensuring the Tech Economy Benefits All of Its Stakeholders*, DATA DIVIDENDS INITIATIVE, https://www.datadividends.org [https://perma.cc/XVB6-B2HT]; *Data as a Property Right*, YANG 2020, https://2020.yang2020.com/policies/data-property-right [https://perma.cc/7SYE-L6ZG]. Legal scholarship and American jurisprudence alike also recognize *de facto* as well as *de jure* rights in data compilations by companies, even if the exact contours of these rights are not always clearly defined. *See, e.g.*, Mary D. Fan, *The Right to Benefit from Big Data as a Public Resource*, 96 N.Y.U. L. REV. 1438, 1466 (2021) ("Applying principles of copyright and trade secret law, courts have recognized property rights in the consumer information that companies compile."); *id.* at 1466 n.154 (collecting cases).

10 Here, the terms "data controllers" and "data subjects" are not used in the statutory sense provided by the European Union's General Data Protection Regulation. *See* Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 4(7), 2016 O.J. (L 119) 1. Rather, they are used in the sense that an ordinary consumer who uses typical online products might understand the terms, them versus us. The "them" refers to the entities that surveil

understands data controllers to be those parties that collect or analyze data about consumers via interactions and observations made when consumers use digital products or services. Data subjects are those consumers about whom observations are recorded as they navigate these digital infrastructures that have become pervasive in everyday life. To date, neither data controllers nor data subjects seem to have received much or any compensation for the value they have produced.

Some stakeholders are taking action to secure their stake in any future use of their data to create AI models. Consider an announcement made in April 2023 by Reddit, a social content aggregation and discussion site, that it would begin charging many AI developers for access to data—that is, for access to the natural language content that Reddit's users (so-called redditors), created while participating in online communities hosted by Reddit.[11] The company has made clear that those seeking high volume access to Reddit data must pay for a premium API service.[12]

As a data source, Reddit is an interesting, even extreme, case as to the question of who should be compensated for data and why. In the case of Reddit, the data collected is content. Thus far, the data that firms have mined from Reddit and used to train LLMs[13] are not simple compilations of facts, but rather collections of expressions of human creativity. Unlike much of the consumer data put into the service of training AI, the content on Reddit is clearly copyright protected not just in aggregate but in individual

---

natural persons and record data about them. The "us" are the natural persons who live their lives in the ordinary way and therefore are likely described in great detail in hundreds if not thousands of databases that record various bits of information about them, sometimes in an identifiable format and sometimes not.

11 *See* Kevin Purdy, *Reddit Will Start Charging AI Models Learning from Its Extremely Human Archives*, ARS TECHNICA (Apr. 19, 2023), https://arstechnica.com/information-technology/2023/04/reddit-will-start-charging-ai-models-learning-from-its-extremely-human-archives [https://perma.cc/3AAV-X4VC] (describing Reddit's plan to charge companies for access to its API while keeping it free for non-commercial projects).

12 *Creating a Healthy Ecosystem for Reddit Data and Reddit Data API Access*, REDDIT (Apr. 18, 2023), https://www.redditinc.com/blog/2023apiupdates [https://perma.cc/7P78-CY3P] [hereinafter REDDIT].

13 Some may believe that the example of LLMs undercut one of the principal observations in this piece, namely that sometimes a far smaller amount of data will do where big data is thought necessary. Firstly, this piece offers this observation not as a universal truth, but as a sufficiently common phenomenon so as to be relevant to policy, even if LLMs are an example where more data may very well always make better LLMs. However, it need not necessarily be the case, and is certainly not proven, that more data always improves LLMs. Indeed, a great deal of care and attention went into preparing the data sets that OpenAI has used to train its models; they have taken an approach that is very far from sucking all potential language information on the internet. *See, e.g.*, Lex Fridman, *Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI*, YOUTUBE (Mar. 25, 2023), https://www.youtube.com/watch?v=L_Guz73e6fw [https://perma.cc/6BDZ-GKGJ] (discussing briefly the process by which OpenAI undertook data curation).

contributions.[14] The work of creating that copyright-protected content was not directly accomplished by Reddit or its employees but rather by people who posted or responded to content on the platform.[15] Reddit even opined, while announcing its new premium API, that "user content is owned by redditors that have created and submitted content on Reddit and cannot be used without permission."[16]

But Reddit has contributed its own labor to the creation of its data stockpiles as well. Redditors would not post their content absent Reddit's technical and business development efforts. Further, that content would not have been available in the useful, easily accessible format enjoyed by AI model developers without Reddit's contributions. In addition to paying for the computing resources to host the content,[17] Reddit created all that goes into attracting and keeping redditors at its site: a platform culture with healthy content moderation practices and norms of civility,[18] good engineering to keep the site consistently accessible,[19] and social engineering

---

14 *See* Uri Y. Hacohen, Amit Elazari & Talia Schwartz-Maor, *A Penny for Their Creations—Apprising Users' Value of Copyrights in Their Social Media Content*, 36 BERKELEY TECH. L.J. 511, 526–27 (2021) (explaining that a substantial portion of the user-generated content on social media will satisfy statutory requirements of originality and fixation to receive protection).

15 Some postings on Reddit are labeled as bot-generated and others are suspected as such. However, Reddit is generally regarded as a source of high-quality human-generated content.

16 REDDIT, *supra* note 12.

17 Reddit is a privately held company and therefore it is particularly difficult to find public information about its operating costs. Nonetheless, if one infers that its costs may be comparable to Twitter, another content sharing platform with a somewhat similarly sized userbase, we can infer that the operating costs should be within one order of magnitude. Consider that Twitter's operating costs in 2022 were reported to be around $1.5 billion for servers alone. *See* Andrew Pantyukhin (@pandrewhk), TWITTER (Dec. 21, 2022, 4:25 AM), https://twitter.com/pandrewhk/status/1605494601673891840 [https://perma.cc/YWJ6-RC2N]. Twitter was recently described as having around 330 million monthly active users while Reddit was described in the same source as having around 430 monthly active users. *See* Kim Cunningham, *Twitter vs. Reddit: Differences, and Which One Is Better*, HISTORY-COMPUTER (Aug. 21, 2023), https://history-computer.com/twitter-vs-reddit [https://perma.cc/JV8G-N5NN].

18 *See* Quina Baterna, *7 Reasons Why Reddit Is the Best Social Media Platform Around*, MAKE USE OF (Mar. 27, 2023), https://www.makeuseof.com/reasons-reddit-best-social-media-platform [https://perma.cc/AQ7K-S5DH] (describing how Reddit moderators create rules for specific communities and remove toxic posts).

19 Again, in contrast to Reddit, Twitter has experienced multiple service outages in the past year, with the press suggesting that this rise in service outages is related to the takeover by Elon Musk. *See* Alex Hern & Dan Milmo, *Rise in Twitter Outages Since Musk Takeover Hints at More Systematic Problems*, THE GUARDIAN (Mar. 8, 2023), https://www.theguardian.com/technology/2023/mar/08/spike-in-twitter-outages-since-musk-takeover-hint-at-more-systemic-problems [https://perma.cc/M9LS-FYED]. From the absence of similar problems, one can infer a reasonable investment in resources to keep a site up and running. *Cf.* Cody Slingerland, *How Much Does Twitter Spend on AWS and Google Cloud?*, CLOUDZERO (Aug. 15, 2023), https://www.cloudzero.com/blog/twitter-aws/ [https://perma.cc/UFQ3-4HPA] ("Musk tweeted that the 2022 cloud infrastructure budget cuts were on the way to reducing non-debt expenses from $4.5 billion to $1.5 billion in 2023.").

methods to keep the site largely limited to human-to-human interaction.[20]

Reddit's data has been used to train modern LLMs, including OpenAI's GPT-3.[21] Reddit's premium API could create a new revenue stream for the firm, potentially one that is quite remunerative. Reddit is seeking a payout for its role in facilitating access to a trove of content useful for training LLMs, but it does not pretend to have a sole proprietary right to the information its premium service will provide. If paying to access data relates to recognizing a copyright interest in that data, one might expect that Reddit would present itself as collecting on behalf of the copyright holders and negotiating usage rights.[22] But this is not how Reddit presents its premium access: It presents itself as providing access to data, not licenses to copyrighted material.[23]

If Reddit's premium access were to prove successful, one could imagine that redditors would come to resent Reddit's de facto monetization of their content. Those redditors might, in turn, pursue the same path Reddit is now exploring. Perhaps the users would bring suit alleging infringement of their copyrights.[24] Or perhaps the redditors would rely more heavily on organic, direct appeals of morality to the company by asking for a fair share. Perhaps they would have some success and come to an agreement with Reddit to share profits. How they might arrive at an agreement as to what portion of the proceeds fairly belonged to each party is an open question, although it clearly has analogs with other instances of divisions between holders of

---

20 *See* Baterna, *supra* note 18 (describing how Reddit makes it easy to interact with strangers all over the world on shared interests and how Karma is awarded based on how other redditors react to your posts and comments in a way that is hard to fake).

21 Many observers assumed that Reddit data had been used to train GPT-3, but a Washington Post investigation confirmed it earlier this year. *See* Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, WASH. POST (Apr. 19, 2023, 6:00 AM), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/?itid=ob_checkout_digital [https://perma.cc/97F2-7WNL].

22 The music industry could be a valuable source of received wisdom and industry best practices were Reddit to embark on the business line of aggregating and distributing copyright licenses. *See, e.g.*, Mandeep Sihota, *A Primer for Valuation of Music Catalogs*, LAW.COM (Sept. 1, 2014), https://www.lawjournalnewsletters.com/sites/lawjournalnewsletters/2014/09/01/a-primer-for-valuation-of-music-catalogs/?slreturn=20230328143635 [https://perma.cc/C4LD-TYWL].

23 A requirement that will apply beginning in June 2023 for use of the Reddit Data API is that API users will comply with any usage requirements or restrictions imposed by the content creators. *See Data API Terms*, REDDIT (Apr. 18, 2023), https://www.redditinc.com/policies/data-api-terms [https://perma.cc/KG4Z-PW8C] ("[N]o other rights or licenses are granted or implied, including any right to use User Content for other purposes, such as for training a machine learning or AI model, without the express permission of rightsholders in the applicable User Content.").

24 Earlier this year, the creator of a popular Reddit community, WallStreetBets, sued Reddit for blocking his attempt to register the term for trademark. Aimee Picchi, *Jamie Rogozinski Created Reddit's Popular WallStreetBets. Now He's Suing over "Nightmare" Ban*, CBS NEWS (Feb. 16, 2023, 6:22 PM), https://www.cbsnews.com/news/reddit-wallstreetbets-founder-jaimie-rogozinski-suing-over-ban-copyright [https://perma.cc/T4FK-USF3]. Reddit itself had previously filed four WallStreetBets-related applications with the U.S. Patent and Trademark Office. *Id.*

capital (in this case, Reddit holding the computing resources and network effects of its large user base) and laborers contributing value[25] through use of that capital (in this case, the redditors who make use of Reddit's computing resources and large user base to enable the production of creative content).[26]

Let's continue the thought experiment further and imagine that the redditors come to an agreement with Reddit in which the company agrees to give them some portion of the premium API proceeds. The negotiations would likely not end there. It would then be up to the redditors to determine how to divide up their portion amongst themselves. Likely, redditors are not all equal in how much value they contribute to Reddit's data stores.[27] Power users[28] could point out that they had been members of Reddit for longer, or had contributed more often or more copiously, or had received more

---

25 Of course, notions of what constitutes value or contribution of value are highly contested. Among a host of notions, economists might point to notions of use value or exchange value. On the other hand, with regard to personal data, privacy law scholars might point to less quantitative notions of value, such as deontologically-grounded dignitarian norms that would reject the exercise of valuation altogether. I think it is possible to avoid debates as to the most descriptive or most normatively defensible definition of value and simply recognize that for practical purposes, we can consider as clearly of interest value related to innovation as well as additional commercial value, acknowledging that these two forms of value creation are not equally normatively compelling. For simplification, the reader can even take the argument in a narrower form of understanding value creation as innovation, as reflected in both the categories of "social innovation" and "market innovation". *See* Richard B. Stewart, *Regulation, Innovation, and Administrative Law: A Conceptual Framework*, 69 CALIF. L. REV. 1256, 1260–61 (1981). *But see* Yafit Lev-Aretz & Katherine J. Strandburg, *Regulation and Innovation: Approaching Market Failure from Both Sides*, 38 YALE J. ON REG. BULL. 1, 6–7 (2020) (arguing that the distinction between the two forms of innovation is confusing and potentially misleading).

26 In this case, the redditors might look to historical indicators of the economic division of profits between labor and capital returns as one indicator of a fair starting point. However, as many labor economists have pointed out, that proportion in technology companies has deviated from historical trends to the detriment of the labor share. *See generally* David Autor, David Dorn, Lawrence F. Katz, Christina Patterson & John Van Reenen, *The Fall of the Labor Share and the Rise of Superstar Firms*, 135 Q.J. ECON. 645 (2020). Likewise, if we are to conceive of the redditors as laborers in an online labor market, we might consider the robust and substantial findings of monopsony in online task environments. *See* Arindrajit Dube, Jeff Jacobs, Suresh Naidu & Siddharth Suri, *Monopsony in Online Labor Markets* 3–4 (Nat'l Bureau of Econ. Rsch., Working Paper No. 24416, 2018).

27 *See, e.g.*, karmarank, *Karma Inequality: 1% of Redditors Have 20% of the Comment Karma*, REDDIT (June 12, 2014), https://www.reddit.com/r/dataisbeautiful/comments/27zyh6/karma_inequality_1_of_redditors_have_20_of_the [https://perma.cc/UPH6-453J?type=image].

28 There is no single definition of a power user, but this term generally designates a natural person who uses a digital product often and in a highly skilled manner. They are valued in industry because they are thought to add value and to drive adoption. *See* Frank L., *How to Identify Your Power Users (And How to Keep Them)*, STREAM (May 31, 2022), https://getstream.io/blog/identify-power-users [https://perma.cc/Z2DH-6D2T].

upvotes.[29] Subreddit moderators—even if they hadn't directly contributed content—might seek special compensation for their role in maintaining the overall high quality of the community; in other words, they might seek compensation for ensuring the high quality of the data even if they did not directly author any data. There would be many morally compelling and seemingly logical arguments as to why different stakeholders deserved more compensation than others.[30] Methods to establish the relative valuation of different contributors, or data subjects, is a topic pursued *infra* in further detail.

But this isn't merely a thought experiment designed to justify dipping into a fascinating body of computer science and economics literature on data valuation or model evaluation. The question of relative value contribution of an increasingly economically valuable and culturally fetishized asset[31] has direct bearing on actively litigated privacy questions. For example, the last stage of our thought experiment, as to how redditors would divide up value amongst themselves, speaks to a process now underway to apportion proceeds from a $750 million settlement, the outcome of litigation against Facebook triggered by the Cambridge Analytica scandal.[32] The settlement funds are to be distributed among American Facebook users who resided in the United States and had a Facebook account during any portion of time between 2007 and 2022.[33] The per-claimant amount will be computed according to the portion of that timeframe during which a claimant had an undeleted Facebook account, and it will likely be tiny in any case given the enormous number of people who meet the inclusion criteria.[34] However, it's

---

29 From the specific point of view of training LLMs, one could also imagine that redditors writing on especially informative/commercially important topics, or writing with higher quality language, might argue that, on its face, their contributions are worth more than the average Reddit contribution.

30 A market-based mechanism could simply be that Reddit paid individual redditors according to how often their particular content passed through the premium API. Depending, however, on how the API was programmed to return content, this might simply mean that Reddit had already made an ex ante decision as to which content it would push most in its premium API access, which would likely be some proxy for which content Reddit judged as most valuable.

31 This is of course just one way to understand data, but it is the aspect of data of interest in this work.

32 *See* Plaintiffs' Notice of Motion and Motion to Certify a Settlement Class and Grant Preliminary Settlement Approval at 14, In Re: Facebook, Inc. Consumer Privacy User Profile Litig., 402 F.Supp.3d 767 (N.D. Cal. 2022) (No. 18-md-02843-VC) ("The Net Settlement Fund will be used to compensate Settlement Class Members for the harms they suffered as a result of Facebook's alleged wrongdoing. It will be allocated to each Authorized Claimant who submits a claim by the Claims Submission Deadline . . . .").

33 *Id.* at 14.

34 *Id.* Even if the full $750 million were allocated directly to claimants—which it will not be—and even if only 20 million claimants file (which seems to be about 10% of a plausible size for the proposed class), this would come to less than $40 per claimant, which seems undervalued given the

not clear that the values should be tiny for everyone; it seems, rather, that some people may truly have suffered enormous harm and others may have just a small amount. One potential proxy for measuring that harm could reasonably be the relative value of their data.[35]

Of course, in the case of distributing funds from a settlement, the funds are intended to compensate for harm, not to pay for value. Nonetheless, just as the value of one data subject's contribution to a data set may not be equal to another's, the harm—even in expectation—suffered by two Facebook users whose accounts were active for equal lengths of time seem unlikely to be equal. One Facebook user might have created an account and failed to delete it but otherwise remained inactive; the other may have logged in many times each day, establishing a dense network of social connections and an extensive history of interactions or personal photographs.

As with redditors, not all Facebook users are equal in their vulnerability to privacy harms or in their contribution of economic value to Facebook's data stores. The same phenomenon of heterogeneity of data value (and of potential harm[36]) is reflected among users of Facebook as among redditors, although in the case of the Facebook data far less of it would seem obviously amenable to copyright protection by the data subjects since the data is largely *not* directly authored[37] by Facebook users and would likely be regarded individually as uncopyrightable facts.

Facebook too has contributed to the generation of its data stockpiles, perhaps as much or more as Reddit has its own. What value does Facebook

---

degree of outrage the American public was said to feel at the time the Cambridge Analytica data leak became widely known. *See, e.g.*, Herb Weisbaum, *Trust in Facebook Has Dropped by 66 Percent Since the Cambridge Analytica Scandal*, NBC NEWS (Apr. 18, 2018, 3:08 PM), https://www.nbcnews.com/business/consumer/trust-facebook-has-dropped-51-percent-cambridge-analytica-scandal-n867011 [https://perma.cc/B3VQ-8MV6].

35 It is of course somewhat more complicated. Data valuation is typically undertaken relative to a specific task. Therefore, to justify this allocation method, those managing the claims would have to define a task, or a set of tasks, in relation to which the data valuation would be calculated. This would, however, have the added benefit of more clearly defining exactly the theory of harm, as that theory of harm could be defined relative to the most egregious risks that had been imposed on data subjects through the data breach. *See infra* Part I.

36 The value of the data contributed and the value of the potential harm from inappropriate disclosure of data need not be highly correlated, but both can be highly heterogeneous among data subjects.

37 I exclude, of course, the case of natural language content on Facebook. In existing publications on the use of Facebook's data, this natural language content has not been emphasized as important either in the company's own publications or in speculation about algorithmic targeting on Facebook, with far more emphasis placed on patterns of likes and clicks. *See e.g.*, José González Cabañas, Ángel Cuevas & Rubén Cuevas, *FDVT: Data Valuation Tool for Facebook Users*, PROC. 2017 CHI CONF. ON HUM. FACTORS IN COMPUTING SYS. 3799, 3801 (2017) (presenting a tool to help users understand the value of their data, but the tool being limited to looking at the number of ads displayed to users and whether they clicked on those ads). Further, for purposes here, the purpose is to contrast direct inputs and "metadata" (data about online behaviors and social relationships) with human-author creative natural language content.

contribute to its massive data sets through its ingenuity in attracting and retaining users as well as its creative and innovative efforts to shape the ways those users interact with one another? What value do the users themselves generate, and for which portion of it ought they to be rewarded? Should it matter whether users are simply "off-gassing"[38] data as they make use of a social media platform, or whether there is skill and labor involved as a necessary precondition to giving meaning to the task of dividing value contributions?[39] Even if some users are inherently "better" than others in the sense of "off-gassing" data that is valuable for a particular purpose, shouldn't they be compensated for their natural talents much as lucky landowners are compensated when by happy chance they have a stake in what turns out to be a valuable deposit of a natural resource?

There is a widely shared but unspoken implication in discussions of privacy law that entities performing surveillance in commerical digital product environments contribute little in the way of skill or innovation beyond the bare practice of recording as much as is economically and legally feasible. Or, alternatively, that any skill, effort, or innovation applied by these entities is unworthy of consideration in policy discussions. For example, many calls for greater data portability and data access rights appear to be founded in part on the assumption that the main source of value creation by a data controller is that of aggregating data, as though it is naturally occurring, or otherwise collecting it in ripe and ready form, like oil or corn. These calls, and related proposals, sometimes include commodity-style grading to distinguish between data sets of relatively higher or lower quality grades. In this case, data recording is sometimes understood, and has sometimes been legally treated, as the capture of a naturally occurring resource, *ferae naturae*.[40] A parallel conceptual framework for evaluating

---

38 Off-gassing is the effortless production of a natural resource that itself took on effort to produce.

39 The history of what constitutes content versus metadata is long and contested, and calls to mind similar debates in litigation and scholarship about the Electronic Communications Privacy Act of 1986 (ECPA) distinction between the content of a communication and what is record or envelope data. *See, e.g.*, Matthew J. Tokson, *The Content/Envelope Distinction in Internet Law*, 50 WM. & MARY L. REV. 2105, 2117–23 (2009) (discussing how courts have applied various parts of the ECPA). The exact dividing line is unimportant, and likewise the examination presented here holds even if these categories are understood to vaguely gesture towards different, contextually-dependent ends of a spectrum rather than clear, naturally occurring categories of information.

40 *See, e.g.*, Vera Bergelson, *It's Personal but Is It Mine? Toward Property Rights in Personal Information*, 37 U.C. DAVIS L. REV. 379, 403 (2003) ("Even though [modern courts] often acknowledge that personal information has become a valuable commodity, they believe that it belongs to no one until collected."); *see also* Fan, *supra* note 9, at 1468–69 (discussing how personal data is akin to the *ferae naturae* pursued by hunters); *Big Data, Big Impact: New Possibilities for International Development*, WORLD ECON. F. (Jan. 22, 2012), http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf [https://perma.cc/WAS4-XKX5]; Kenneth Cukier, *Data, Data Everywhere*, ECONOMIST (Feb. 27,

data capture is that the data is *our data*, us being human society or individual data subjects. In this narrative, data capture is an imposition on us, either as a society or as individual data subjects. In this case, control rights such as statutory rights to data access or data deletion are portrayed as protections for the right of personality or for the fundamental human right of dignity.[41] This work looks to complicate the discussion by emphasizing a point that has not been much considered in the literature: the exercise of skill—heterogeneous skill—by data controllers and data processors.

I take the lack of discussion in the literature as to the relative valuation of data subjects' contributions and data processors'[42] contributions as suggesting either a silent consensus that any such variation is unimportant or as evidence of an undertheorized empirical fact in the Big Data economy. In this essay, I present evidence that both data subjects and data controllers exhibit significant variations in the measured value of their contributions to the standard Big Data pipeline. I then establish that such variations are worth considering in technology policy for privacy, competition, and innovation.

The essay proceeds as follows. In Part I, I present a discussion of three data valuation methods—willingness to pay, willingness to accept, and model performance differentials—to establish the relative value of different kinds of data. Under each method, simple valuation exercises show a range of relative valuations attributed to different individual or groups of data subjects, revealing that it is not difficult to discover variations in data subject value in real world data sets. In Part II, I present a discussion of variations in model performance that highlight the importance of the skill of the technician who develops the model, but also help explain the difference in the value that a data set, as compared to a technician, contributes in solving a particular problem. This establishes that heterogeneity in relative data subject valuations is mirrored by heterogeneity of machine learning (ML) operator skill and in heterogeneity of the relative contributions of data and ML operators in solving a given task.

Following this descriptive work, in Part III I present potential

---

2010), https://www.economist.com/special-report/2010/02/27/data-data-everywhere [https://perma.cc/XP3N-QXMS] ("Data are becoming the new raw material of business: an economic input almost on a par with capital and labour.").

41 *See, e.g.*, Gabriela Zanfir, *The Right to Data Portability in the Context of the EU Data Protection Reform*, 2 INT'L DATA PRIV. L., no. 3, 2012, at 3 (arguing that, considering that the amount of data collected on individuals can create their digital personalities, the thought "of data collectors forbidding [them] to transfer their stack of data from one service provider to another could seem to be a violation of human rights").

42 I use terms associated with data protection law as shorthand that I expect to be commonly understood but that need not be strictly confined to any precise statutory definition. Further, I use terms referring to data controllers and data processors largely interchangeably and in opposition to data subjects. In short, I mean to refer to those about whom data is collected as opposed to those who collect and use data for commercial purposes.

implications for privacy, competition, and innovation. In the privacy law discussion, ignoring heterogeneity means that we miss opportunities to reduce data flow and data collections, and to allow for greater specificity in theories of privacy harms associated with personal data. In competition law, ignoring heterogeneity means that we have likely put too little thought into how data portability and interoperability ought to work and what incentives might be produced in data-driven industries when firms know they will face data sharing obligations. In the innovation space, we are likely to see too much investment in the wrong kinds of data infrastructure and too little incentivization of skilled operation of either data production or data analysis given the widespread assumption that such skill is less important than access to raw data. We are likely underinvesting in the most innovative data-driven technologies by taking data as a given rather than a skill and a valuable, heterogeneous asset. The overall lessons drawn here focus on the domain of commercial surveillance and consumer privacy, but future work should consider how these same observations might guide other uses of big data and algorithmic development.

I
DATA SUBJECT HETEROGENEITY

Industry practitioners and computational researchers commonly use three methods to determine the value created by a group of data subjects within a data set. Two of these categories, namely willingness to pay (WTP)[43] and willingness to accept (WTA)[44] are used both in industry and also in behavioral economics research.[45] The third category of valuation is computational methods, in which calculations for an individual data point or a subset of data are made with respect to a particular task and conditional on other data available for that task. Computational methods do not appear to be typical in industry, likely due to their computational cost, but they have been proposed as a more accurate and fair method of attributing value creation to data subjects as compared to WTP or WTA.

There is early-stage literature surveying ways to value data, but there is no dominant or all-purpose method. A recent survey of data valuation

43 That is, what an entity not in possession of the data is willing to pay to access that data.

44 That is, what an entity in possession of the data is willing to accept to make that data available to another. Because data is a non-rival good, implausible scenarios of a full data transfer are not contemplated here.

45 In other work, I have put forward concerns as to what market measures of privacy preferences might indicate. However, the point as to heterogeneity of valuation remains even if I have concerns as to a large bias in the valuations due to the taboo nature of trading privacy for money. *See* Aileen Nielsen, *Taboo and Technology: Experimental Studies of Data Protection Reform*, N.Y.U. J. LEGIS. & PUB. POL'Y (forthcoming 2024) (manuscript at 54–59), https://docs.google.com/document/d/1zl3EOFL3tq6QXAFfw6vUtpIeLbI0b1IEYkRwk_oDf3A/edit#heading=h.6hvf411pjszs [https://perma.cc/5FX5-H9KV].

methods in the *Harvard Data Science Review*—the only categorization of such methods identified in a recent literature review—proposed a framework that involved three prongs: market-based models, economic models, and dimensional models.[46] In contrast to that survey, this work considers both market-based and economic models as one category of methods that ultimately rely on WTP or WTA. Likewise, dimensional models reflect only a subset of the computational work discussed here. This paper does not purport to fully survey existing work on data valuation. Instead, it demonstrates that data subject heterogeneity is common and substantial.

A. Market Measures

Market measures convert subsets of a data set, or even individual data points, to monetary valuations through elicitation of WTP or WTA. Such metrics are found both in industry and in academic research. In industry, most publicly available data valuations and likely most business-to-business transactions are quoted for access to data-driven services rather than to data itself. In other words, there is far more selling of data-driven services than data directly.[47] This paper focuses on market measures not only due to their dominance in business transactions but also because the link between consumer data and consumer attention is the dominant business model that drives commerce in the Big Data economy.[48]

1. Willingness to Pay

Consider pricing data available from two of the largest companies[49] in the world, Alphabet (the owner of Google) and Meta (the owner of Facebook). Both companies are substantial presences in the market for human attention and likely keep some of the world's largest stores of

---

46 *See* Mike Fleckenstein, Ali Obaida & Nektaria Tryfona, *A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model*, HARV. DATA SCI. REV., Winter 2023, at 2–4, https://hdsr.mitpress.mit.edu/pub/1qxkrnig/release/1 [https://perma.cc/JW2C-MED3] ("We acknowledge that no single approach to data valuation exists today, and that different approaches—even a combination of approaches—can be used, depending on the use case.").

47 Financial Management Association International, *Valuing Data as an Asset – Laura Veldkamp*, YOUTUBE (Feb. 22, 2023), https://www.youtube.com/watch?v=l_oDLKha4co [https://perma.cc/V6XF-D6ZQ].

48 *See, e.g.*, Daniel L. Rubinfeld & Michal S. Gal, *Access Barriers to Big Data*, 59 ARIZ. L. REV. 339, 342, 359 (2017) (discussing how data quality affects advertisers, users, and others); *see also* Elettra Bietti, The Structure of Consumer Choice: Integrating Antitrust and Utilities in Digital Platform Markets 7 (Aug. 2022) (unpublished manuscript) (on file with author).

49 In this case, the size of a company is defined by market capitalization in 2023. *See List of Public Corporations by Market Capitalization*, WIKIPEDIA, https://en.wikipedia.org/wiki/List_of_public_corporations_by_market_capitalization [https://perma.cc/R9V5-3N9G] (last visited Oct. 30, 2023) (listing the top ten publicly traded companies by market capitalization in 2023, which includes Alphabet and Meta).

consumer data.[50] There is some data available regarding WTP for data-driven services from these companies, particularly in the case of attention-related data-driven services for behaviorally-targeted advertising shown on Google and Facebook. While data markets do not typically run on an auction system,[51] ad impressions, including those on Google and Facebook, largely run on a real-time auctioning system, wherein individual advertisers go through automated models to bid for specific segments of user attention.[52] Because in an auction system the WTP must meet or exceed the WTA, the results of such an auction provide evidence of advertisers' WTP for attention from heterogeneous groups.

The cost of displaying a successful advertisement varies, and variations in cost correlate with data subject characteristics. These correlations can reveal advertisers' WTP for clicks or impressions from particular user populations.[53] For example, a 2012 study found that the cost per thousand impressions and cost per click were higher for Facebook ads targeted toward women than for ads targeted toward men.[54] More recently, Google updated its AdWords advertising service to allow demographic targeting: Google AdWords now allows advertisers to create bid adjustments for different demographic groups, thereby facilitating heterogeneity in costs for impressions and clicks from varied age and gender groups.[55]

---

50 How one determines the size of a data set is not necessarily a simple problem, much like that of valuation. For example, how does one compare a data set with more individual data points but fewer attributes as compared to a data set with fewer individual data points but many more recorded attributes? What constitutes a large data set will vary over time, by industry, and according to the task one is seeking to accomplish. There is no official metric.

51 This is likely due to the non-rivalry of data, among other reasons.

52 *See About the Ad Auction*, GOOGLE ADSENSE HELP, https://support.google.com/adsense/answer/160525?hl=en#:~:text=AdSense%20uses%20an%20auction%20to,user%20sees%20on%20your%20site [https://perma.cc/869J-NZ6K] (last visited Nov. 13, 2023) (describing the ad auction system utilized by Google to determine which ads are shown on what sites); *The Ad Auction Explained*, META ADS, https://www.facebook.com/business/ads/ad-auction [https://perma.cc/2KMZ-CYH2] (last visited Nov. 13, 2023) (explaining that which ads are shown to which populations on Facebook is determined by a multifactor auction process).

53 A further question not addressed here is who benefits from this value, as between the data subject/ad viewer, the buyer of the ad, and the platform showing the ad.

54 *Women vs. Men: Guess Who's More Likely to Click on a Facebook Ad?*, THE REALTIME REPORT (Sept. 26, 2012), https://therealtimereport.com/2012/09/26/women-vs-men-guess-whos-more-likely-to-click-on-a-facebook-ad [https://perma.cc/52B2-45KL]; *see also Social Media Insights: Men Are Cheap*, SDRS CREATIVE, http://sdrscreative.com/works/kenshoo-social-men-are-cheap [https://perma.cc/Y7J9-VW9K] (last visited Nov. 13, 2023) (showing infographics comparing cost per thousand impressions and cost per click for Facebook ads by gender).

55 *See* Mark Irvine, *New in AdWords: Demographic Targeting for Search Campaigns*, WORDSTREAM (Nov. 24, 2022), https://www.wordstream.com/blog/ws/2016/09/12/demographic-targeting-for-search-campaigns [https://perma.cc/YZ6M-G6WT] (describing the new demographic targeting functions available to advertisers in Google AdWords search campaigns, including ability to make bid adjustments and see a breakdown of cost by age and gender groups).

Of course, the fee a firm charges to display a one-time advertising to a particular user in a particular segment of the user population is not necessarily determinative of the value the firm ascribes to that user. The average man who uses Facebook apparently commands a lower cost per click than does the average woman, but this might be partly explained by a greater willingness of men to click on the ads shown to them.[56] The average man might well earn more revenue for Facebook than does the average woman despite a lower per-click or per-impression valuation and so in fact be valued more highly by the company.[57] Of course, we cannot know the answer to these questions from the information available. What we can infer from the data is that it is improbable that men and women are exactly equally valuable to Facebook for generating ad revenue. Because data collection is a core component of online advertising, this variation may suggest that the data collected about one gender may generate more revenue than that collected about another.[58] Put simply, when data controllers weigh the relative value (to them) of data from various populations of data subjects, they are not likely to find equal value.

Facebook and Google primarily earn revenue by selling access to data-driven products, particularly by selling access to data-driven advertisements. From this, one might infer that Facebook and Google likely do attribute differential values to different segments of their user base, given that firms receive differential WTP for providing advertising access to different segments, and thereby have different degrees of opportunity to monetize different segments. It is possible that data controllers already possess information regarding the heterogeneous monetary valuation of different data subjects based on their experience in marketing data-driven services. Likewise, data controllers likely have their own proprietary valuations of

---

56 SDRS CREATIVE, *supra* note 54.

57 *See id.* (explaining that advertisers pay more money for ads targeted at men on Facebook).

58 This need not imply that the company necessarily weighs the needs or preferences of one gender more than the other. Likewise, this should not be taken to mean that the data of one gender over the other is intrinsically more valuable in some deontological sense. Rather, from the perspective of a company executive evaluating potential economic returns from one gender versus the other, these returns may not be equal. Therefore, these returns may influence an executive to be willing to pay one gender more than the other, say to continue using the platform, if that bargain were put to them. However, it is also worth noting that it *need not be the case* that the company *must not* discriminate against one gender over the other. After all, it is not clear that the company has any legal obligation to treat the genders equally, as it might make the case that it is not a place of public accommodation, much like Uber has attempted to do in the context of compliance with the Americans with Disabilities Act. *See* Carmen Carballo, *Tap a Button, Get Denied: Uber's Noncompliance with the ADA*, MINN. L. REV. DE NOVO BLOG (Apr. 24, 2020), https://libpubsdss.lib.umn.edu/minnesotalawreviewprod/2020/04/24/tap-a-button-get-denied-ubers-noncompliance-with-the-ada [https://perma.cc/L9NZ-Z25Q] (noting that Uber tried to argue that it was not subject to the Americans with Disabilities Act because the company is not a public accommodation as defined in the ADA). The question of whether these platforms are covered by any federal anti-discrimination legislation remains an open question.

their data subjects' relative contributions in ways that will vary across data subjects. A world of scholarship and regulation that is insensitive to these variations in value, even when data controllers themselves are highly informed and highly attuned to this variation, is a world where the potential benefits to the public of such variation will likely be foregone. To preview some benefits, it is possible that substantially less data collection can be undertaken with the same yield on algorithmic performance.[59] Likewise, it is possible that smaller firms can enter data rich markets if they concentrate on obtaining high quality data rather than merely big data.[60] As discussed in Part III, opportunities exist to turn this heterogeneity into gains for social welfare by calibrating privacy, competition, and innovation policy appropriately.

2. Willingness to Accept

Consider next the measure of WTA, which, like WTP, translates a desire into a monetary valuation. WTA asks how much a potential seller values a good based on the price at which she is willing to part with that good. For example, imagine that Facebook charges ten cents per impression for an ad targeted at a particular demographic; then we can say with certainty that Facebook's WTA for that good—in this case the intangible good of an ad impression at a particular time—is no higher than ten cents. In reality, it may even be lower, as the ultimate market price will not only reflect Facebook's WTA, but also its bargaining position given current market conditions.

Industry materials furnish examples of firm WTA in the case of data brokers selling data on open markets.[61] Results of academic research furnish examples of consumer WTA in lab experiments where consumers are offered some monetary value for their personal data.[62] As discussed below, in both data broker materials and academic experiments, heterogeneity in data valuations correlates with various observed group demographic

---

59 *See infra* notes 84–91, 95–105 and accompanying text (discussing various studies indicating that not all data points improve model performance equally).

60 *See infra* notes 84–91, 95–105 and accompanying text (citing research rejecting notion that data set volume is necessarily tied to data set quality).

61 *See infra* notes 67–76 and accompanying text.

62 *See, e.g.*, Avinash Collis, Alex Moehring, Ananya Sen & Alessandro Acquisti, Information Frictions and Heterogeneity in Valuations of Personal Data (Sept. 2023) (unpublished manuscript) (available at http://dx.doi.org/10.2139/ssrn.3974826 [https://perma.cc/LWF3-PSUG]) (investigating heterogeneity in monetary valuation for personal data across different demographic groups before and after being presented with certain information about real-world transactions involving social media data); *see also, e.g.*, Yi-Shan Lee & Roberto A. Weber, Revealed Privacy Preferences: Are Privacy Choices Rational? (2022) (unpublished manuscript) (available at https://static1.squarespace.com/static/58318e41b8a79b98acd4fb9f/t/625343c458497c282010ae6b/1649624007826/Revealed+Privacy+Preferences+2022-03-10.pdf [https://perma.cc/VKC7-AH2G]) (exploring the relationship between individuals' privacy attitudes and their WTA for trading off personal information for money).

characteristics, showing first that there is heterogeneity and second that firms are likely aware of demographic trends predictive of that heterogeneity.

a. Firm WTA

Consider first firms' WTA. Data brokers, also known as data aggregators, have emerged as an entire industry focused on making data available for a monetary price, sometimes publicly announcing a WTA value (that is, the quoted price). These firms make their living off of collecting data from various primary or secondary collectors and aggregating that data into large data sets, before marketing these data sets to the public or to specific kinds of businesses.[63] While much reviled in some groups,[64] and the target of special regulations in some states,[65] data brokers continue to prosper and to provide data at a price.[66] Their public sales of data offer a public record of the relative valuations—based on WTA—that they ascribe to various data sets.

While there does not appear to be an empirical literature on the practices of data brokers,[67] an inspection of public listings—such as those available on websites for data markets or for individual firms' catalogs—reveals that data brokers tend to sell data by content and volume, without considering the relative value of different data points within the data set. Where brokers differentiate between data sets by data quality, they seem exclusively to focus on whether a particular data point "worked," that is, whether a particular individual identified as a business target did in fact lead to business

63 Consider for example the data aggregators at issue in *Sorrell v. IMS Health Inc.*, 564 U.S. 552 (2011). These aggregators purchased information from pharmacists regarding the prescribing records of physicians whose prescriptions had been filled at individual pharmacies. *Id.* at 558. They then aggregated this information across all pharmacies and leased reports on prescriber behavior to pharmaceutical manufacturers, who then used the information for targeted in-person advertising to doctors. *Id.* at 557–58.

64 *See, e.g.*, Data Collaboration Alliance, *I Hate Data Brokers. By the End of This Talk, You Will Too!*, YOUTUBE (Oct. 29, 2022), https://www.youtube.com/watch?v=CEZzhztvqwA [https://perma.cc/N8D6-L4VH] (criticizing data brokers for invading privacy and potentially causing personal safety issues).

65 For example, California and Vermont require annual registration for data brokers. *Data Broker Registration*, ROB BONTA ATT'Y GEN., https://oag.ca.gov/data-broker/register [https://perma.cc/JCY3-SUK7] (last visited Nov. 17, 2023); *Data Brokers*, VT. SEC'Y OF STATE, https://sos.vermont.gov/corporations/other-services/data-brokers [https://perma.cc/P2MZ-QJGE] (last visited Nov. 17, 2023).

66 *See* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY I (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [https://perma.cc/HNP8-TUSX] ("Data brokers—companies that collect consumers' personal information and resell or share that information with others—are important participants in . . . Big Data economy.").

67 Indeed, the empirical literature on data sales is quite limited and in need of expansion. *See* Nielsen, *supra* note 45 (manuscript at 83–84) (reviewing the existing literature).

(or lead to some other measurable success). For example, Gizmodo identified thirty-two data brokers selling data on people labeled as "actively pregnant," "shopping for maternity products," "interested in pregnancy," or "intending to become pregnant."[68] The thirty-two brokers offered a common pricing structure: a flat price per number of users that the data buyer successfully reached with online ad targeting based on the data set.[69] The data was valued, in either case, using a flat count, much like barrels of oil or bushels of corn. Data brokers do not publicly offer pricing on whole data sets that distinguishes among data points; data is instead sold like a raw resource.

Nonetheless, heterogeneity in data set pricing appears when data brokers sell separate data sets that differ by a single attribute. In other words, one can readily identify data sets for sale that contain the same information about different subpopulations; data brokers often sell those sets at different prices. This heterogeneity can be read right off public price lists. Consider for example WTA listings (pricing offers) by AmeriList, a company that sells mailing lists for various demographic groups.[70] The Affluent Seniors Mailing List costs $65/M (per million)[71] while the Gun Owners Mailing List costs $85/M.[72] All else equal, an affluent senior's data point is discounted more than 20% relative to a gun owner's data point.

Partisanship also affects individual data point value. WTA is the same ($85/M) for the data of registered Democratic[73] or Republican[74] voters, but information for political donors is priced differently, at $65/M[75] for Democratic donors and $70/M[76] for Republican donors. Subsets of the same data differ substantially according to characteristics of the data subject. Within just one company, Republican donors' data points are valued more than those of Democratic donors, just as gun owners produce more valuable

---

68 Shoshana Wodinsky & Kyle Barr, *These Companies Know When You're Pregnant—And They're Not Keeping It Secret*, GIZMODO (Aug. 18, 2022, 10:00 AM), https://gizmodo.com/data-brokers-selling-pregnancy-roe-v-wade-abortion-1849148426 [https://perma.cc/86QF-MN37].

69 *Id.*

70 *About Us*, AMERILIST, https://www.amerilist.com/aboutus [https://perma.cc/2E75-J7DP] (last visited Nov. 17, 2023).

71 *Affluent Seniors Mailing List*, AMERILIST, https://www.amerilist.com/affluent-seniors-mailing-list [https://perma.cc/8ZPC-4QE9] (last visited Nov. 17, 2023).

72 *Gun Owners Mailing List*, AMERILIST, https://www.amerilist.com/gun-owners-mailing-list [https://perma.cc/4RKM-YET2] (last visited Nov. 17, 2023).

73 *Registered Democrats – Voters Mailing List*, AMERILIST, https://www.amerilist.com/registered-democrats-voters-mailing-list [https://perma.cc/P2CL-UJTE] (last visited Nov. 17, 2023).

74 *Registered Republicans – Voters Mailing List*, AMERILIST, https://www.amerilist.com/registered-republicans-voters-mailing-list [https://perma.cc/C5Q9-9VAX] (last visited Nov. 17, 2023).

75 *Democrat Donors Mailing List*, AMERILIST, https://www.amerilist.com/democrat-donors-mailing-list [https://perma.cc/RLF6-CVWJ] (last visited Nov. 17, 2023).

76 *Republican Party*, AMERILIST, https://www.amerilist.com/republicans [https://perma.cc/C33E-G9X4] (last visited Nov. 17, 2023).

data points (as measured by WTA) than do affluent seniors.

While this pricing disparity might be understood as a WTA disparity in value between data sets, the information in all of these proposals is the same—only the specific data points differ. One could imagine combining all these data sets and pricing the resulting data set using a weighted average of the combination. Such a price would reflect valuation heterogeneity between individual points.

The WTA price for a particular data set could reflect any number of factors, such as the difficulty of assembling it, the likelihood that contact data will lead to a contact, or even simply vendor preferences. For instance, the price difference between Democratic and Republican political donors could hypothetically stem from a shortage of information about Republican donors as compared to Democratic donors, making it harder to compile a Republican donor list. Alternatively, the price difference could result from the vendor's favoritism towards Democratic candidates.[77] But the cause of variation in WTA is unimportant for purposes of this work. The simple descriptive fact is that distinct groups of data subjects command different prices on the open market. This does not necessarily indicate the true value of the data even for those who believe a true value of data is computable, but it reflects the data broker's investment returns and estimate of their customers' WTP. In any case, it is clear that some identities—that is, certain data points—command higher or lower prices than others.

b. Consumer WTA

Data holds some value beyond its use. We can measure that value by reading individuals' WTA for using or not using their data: This is known as the market for privacy. If we consider the WTA payment that data subjects are willing to receive to give up their data, we can obtain another measure of the data's value. Behavioral economists have found significant differences in the value that data subjects place on their data, even more than what is seen in data broker WTA reviews.

A working paper by Collis, Moehring, Sen, and Acquisti examines participants' WTA in monetary compensation for their Facebook data and

77 This seems unlikely to be the case for any number of reasons, including the fact that most companies are reluctant to be perceived as siding with one political party over the other and because, at least in the case of public companies, a duty to shareholders that would not allow for political subsidies unrelated to returns on investment. *Cf.* Spencer MacColl, *Democrats and Republicans Sharing Big-Dollar Donors, DCCC's Million-Dollar Pay-Off and More in Capital Eye Opener: November 10*, OPENSECRETS (Nov. 10, 2010), https://www.opensecrets.org/news/2010/11/democrats-and-republicans-sharing-b/ [https://perma.cc/23CR-896Z] ("The Democratic Governors Association and Republican Governors Association share 48 top donors, a Center for Responsive Politics analysis of the group's top 100 non-individual donors indicates.").

how participants' valuations for their data are affected by exposure to information about real-world transactions involving social media data.[78] The authors found significant variations in valuations among demographic groups and note that women, Black, and low-income participants were more likely to indicate a lower WTA than other participants—sometimes several orders of magnitude lower.[79] Even after correcting participants' information asymmetry by providing them with market-pricing information of personal data, the authors found that Black and female participants still systematically sought lower WTA bids than other groups.[80] This indicates that consumers from these historically underprivileged groups would likely be disadvantaged in a system where data subjects are encouraged to be active participants in the market for privacy, because they would start with a less ambitious bargaining position. Indeed, in addition to providing information about extremely heterogeneous views on the value of privacy, this research suggests that privacy law must do more than simply correct information asymmetry—in contradiction to laws that would seek to convey more market value information to data subjects as an intervention to make digital environments safer by some notion of that term.[81]

Variations in the willingness of ordinary people to accept money in exchange for data give another lens into the value of personal data—specifically, the value of *not* transmitting data. Heterogeneity in data valuation among data subjects in the study conducted by Collis et al. was extreme, ranging from less than $250 to $10,000 or more.[82] Such variations certainly ought to be accounted for because they potentially reflect quite differential preferences as to privacy and individual control of data.[83]

The foregoing discussion should not be read to imply in any way that people are worth more or less, but rather that data about certain kinds of people will fetch more in the market than will other kinds of data, due either to different asking prices by data brokers or to different estimated monetary value by the data subjects themselves. Whether WTP or WTA are normatively acceptable ways of determining the value of information or of

---

78 *See* Collis et al., *supra* note 62.

79 *See id.* at 13, 33 fig.2, 34 fig.3.

80 *See id.* at 17, 45 fig.A.5.

81 *See, e.g.*, Designing Accounting Safeguards to Help Broaden Oversight and Regulations on Data Act, S. 1951, 116th Cong. § 3(a)(1)(A)(i) (2019) (requiring commercial data providers to routinely "provide each user . . . with an assessment of the economic value that the commercial data operator places on the data of that user").

82 Collis et al., *supra* note 62, at 11–12.

83 There are reasons to question whether typical consumers are able to price their data when invited to do so. A discussion of this literature is beyond the scope of this work, but I acknowledge that such measurements are suspect as indicators of the true value of privacy to a particular individual. *See* Nielsen, *supra* note 45 (manuscript at 54–59) (reviewing the literature). All I posit here is that—whatever WTA for consumers is measuring—it is indeed highly variable.

privacy is a separate question altogether from the descriptive task here of establishing that such variation can be easily found in real world situations.

A topic left for future work and not treated here is the obvious concern about distributive justice and equity in society: If certain segments of the population have data that is systematically judged to be less economically valuable, this adds an additional reason for concern as to whether "data property rights" solutions will ultimately lead to the social outcomes sought by those who propose privacy law reform in general or data property rights in particular. The evidence presented in this work suggests that granting data property rights would not equal distribution of the economic gains of the Big Data economy in a way consistent with the social values of egalitarianism or protection for systematically disadvantaged groups.

B. Model Performance Differentials

A third method of defining data value focuses exclusively on the use of that data to train algorithms. The "Model Performance Differential" measure asks how much a data point or set of data points would enhance the performance of a particular algorithm in a particular task. Researchers refer to the average marginal contribution of a data subset—after all possible combinations of the data have been considered—as the "Shapley value" of the subset.[84] Determining a data point's exact Shapley value is computationally expensive because the data point's contribution to every possible training subset (of all sizes, from one subset to the full data set) must be evaluated in the process.[85] Despite its resource-intensiveness, this method is considered the gold standard for assessing the relative contribution of a particular subset of data to a machine learning model's performance on a particular task.[86]

---

84 The "Shapley value" is named after economist Lloyd Shapley, who first developed the concept in game theory to assess the "value" for an essential, n-person game. *See* L. S. SHAPLEY, U.S. AIR FORCE PROJECT RAND, NOTES ON THE N-PERSON GAME — II: THE VALUE OF AN N-PERSON GAME (1951), https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.pdf [https://perma.cc/Z84S-WJJ2].

85 Amirata Ghorbani & James Zou, *Data Shapley: Equitable Valuation of Data for Machine Learning*, 97 PROC. OF THE 36TH INT'L CONF. ON MACH. LEARNING 2242, 2244 (2019), https://proceedings.mlr.press/v97/ghorbani19c.html [https://perma.cc/CJC6-V8VZ].

86 *See id.* at 2243 (noting that the Shapley value is the only measure that "satisfies three natural properties of equitable [data] valuation," including the possibility of a null value for data that does not change performance, symmetry in value for data with equal contribution to performance, and composability to allow computation of the sum of values across multiple subtasks). Shapley values have also been heavily used in the "explainable AI" literature. *See, e.g.*, *An Introduction to Explainable AI with Shapley Values*, SHAP, https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html [https://perma.cc/2TAG-E5LF] (last visited Nov. 17, 2023) (discussing how to explain machine learning models with Shapley values).

In recent years, substantial work has focused on approximation methods to estimate data Shapley values so that they can be determined with less computational cost. In a seminal paper, computer scientists Amirata Ghorbani and James Zou showed that Shapley values can be estimated to allow for real world applications, and further that the utility of such estimations was substantial, even in real world data sets.[87] Ghorbani and Zou used data sets of cancer patients from twenty-two health centers located across the United Kingdom to build models to predict cancer.[88] Computing an estimated Shapley value for each data set, they identified that patient data coming from one particular location, Nottingham, reduced the performance of a model to predict colon cancer.[89] Further investigation showed that this was because the data from Nottingham reflected a distinct population of patients with a different distribution of characteristics relative to the national profile.[90] Inclusion of Nottingham data resulted in worse performance for the national modeling task at interest. This use of estimated Shapley values clearly shows that bigger data is not always better—for some purposes, certain data points are worth nothing or less than nothing.

Ghorbani and Zou's work is not only useful for identifying which portions of data ought to be removed as unhelpful, but also which portions of data will contribute to improving a model. In Figure 1, reproduced from Ghorbani and Zou's paper, the authors showed that mindless use of big data—simply adding more randomly chosen data—resulted in diminishing returns to model performance, even as adding data points with positive estimated Shapley values continued to improve model performance.[91] The plot shows in stark terms that not all data are created equal.[92]

---

87 *See* AMIRATA GHORBANI & JAMES Y. ZOU, WHAT IS YOUR DATA WORTH? EQUITABLE VALUATION OF DATA (2019), https://arxiv.org/pdf/1904.02868.pdf [https://perma.cc/6ESR-Z2L8] (estimating Shapley values of various real world data sets to assess their value as defined by their ability to improve predictive model performance).

88 *Id.* at 7.

89 *Id.*

90 *Id.* (noting that age was the strongest factor predictive of colon cancer for the entire population, but in Nottingham there was "no significant distributional difference between the age of healthy and diagnosed patients").

91 *Id.* at 6, 8 fig.3(c).

92 Consider as an analogy the received wisdom of the "10x" employee, who is able to bring significantly greater value to a company than the "average" employee. Lucy Brewster, *Why Recruiting '10x' Employees Remains So Difficult Even as Employers Start to Win Back the Upper Hand in Hiring*, YAHOO! FIN. (Apr. 6, 2023), https://finance.yahoo.com/news/why-recruiting-10x-employees-remains-104928899.html [https://perma.cc/N49S-MUU6]. Just as not all employees are created equal, not all data are created equal.

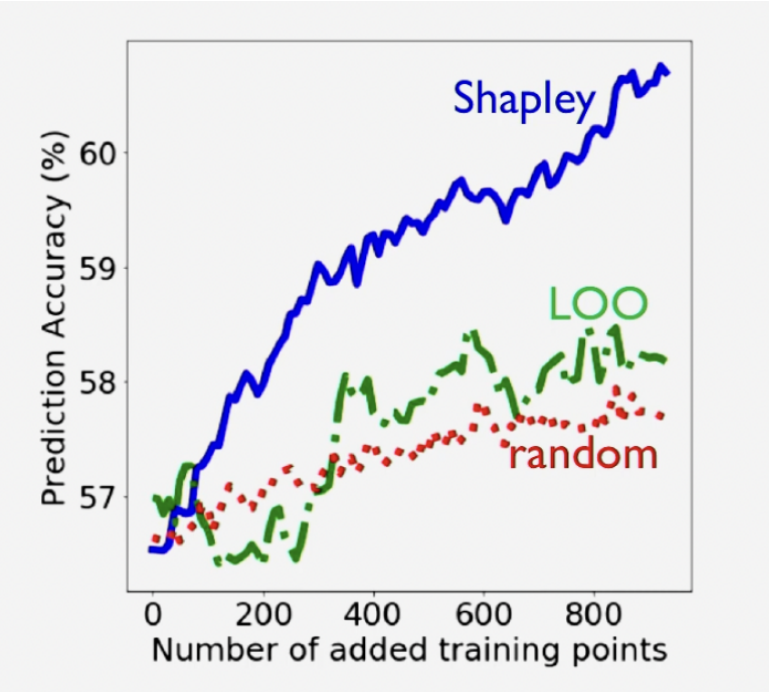

Figure 1: Plot reproduced from Ghorbani and Zou in which the authors showed how choosing data points with positive estimated Shapley values (blue) led to dramatically faster improvement in model performance than did randomly selected data points (red).[93]

Indeed, one proxy that was accepted as some measure of a data set's value was its size, or volume, relative to other data sets.[94] However, Ghorbani and Zou's work gave the lie to the notion that the size of data is an adequate measure of its value.

---

93 GHORBANI & ZOU, *supra* note 87, at 8 fig.3(c). Green represents "leave one out" (LOO) values, which assess the relative value of data points by examining the difference in model performance with a single data point removed and all else remaining equal. Santiago Andrés Azcoitia, Marius Paraschiv & Nikolaos Laoutaris, *Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces*, PROC. 30TH INT'L CONF. ON ADVANCES IN GEOGRAPHIC INFO. SYS., Nov. 2022, at 4 https://doi.org/10.1145/3557915.3561470 [https://perma.cc/K8FC-MKM9]. Until recently, LOO was believed to be a reasonable proxy for the Shapley value. *See id.* at 2 (listing LOO as one alternative heuristic used to estimate data value). The authors showed in this work, however, that LOO is no better than a random selection in determining the value of a data point to model performance. This finding is backed up by subsequent research.

94 *See* Santiago Andrés Azcoitia, Marius Paraschiv & Nikolaos Laoutaris, *Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces*, PROC. 30TH INT'L CONF. ON ADVANCES IN GEOGRAPHIC INFO. SYS., Nov. 2022, at 2, https://doi.org/10.1145/3557915.3561470 [https://perma.cc/K8FC-MKM9] (listing data volume as one alternative heuristic to estimate data value).

The health data Ghorbani and Zou examined in their study is in no way atypical. Indeed, other studies provide even more startling showings that the value of big data is questionable and that the relative spread in value contributed by different subsets of data or individual data subjects can be enormous. In a 2022 study by Azcoitia, Paraschiv, and Laoutaris, the authors studied the relative value of data from different taxi companies in predicting future taxi demand in two large cities, Chicago and New York.[95] They found that sufficiently large companies held enough data to be able to *individually* predict the reference city's overall demand with over 96% accuracy.[96] In Chicago, individual data sets from the city's fifteen largest taxi companies, as well as an aggregate data set combining data from the remaining taxi companies (which account for less than 5% of total taxi demand), could each predict citywide demand with accuracy similar to using all data from all companies: The overall spread in accuracy was small, ranging from over 96% to over 98%.[97] Thus, performance of the predictive model was not purely dependent on the volume of data, and using certain subsets of data could achieve almost identical model performance as using the entire data set.[98] Only the Chicago results are reported here, but the conclusions were globally similar for the Chicago and New York data sets.[99]

The authors also studied the ability of the individual data subsets from each taxi company to predict per-district demand.[100] This was a more challenging problem because, as the subsets became smaller, variations in their usefulness for prediction grew.[101] The researchers found that in the case of many districts with smaller volumes of taxi ride data available, some improvement in predictions could be achieved by combining data from two or more taxi companies.[102] However, even in these cases, the relative value contributed by different taxi companies to prediction accuracy varied widely, including sometimes negative contributions or differences in contribution by several orders of magnitude even among those companies with positive data Shapley values.[103] In other words, not all taxi cab companies' data were created equal for this task: Some companies' data were *orders of magnitude*

95 *Id.* at 4, 7.

96 *Id*. at 2.

97 *See id.* at 4, 4 tbl.1 (describing city-wide demand prediction accuracy rates generated by seventeen distinct data sets: one each from the city's fifteen main taxi companies, one from a hypothetical company that is an aggregation of all remaining taxi companies, and one that aggregates all data across all taxi companies).

98 *See id.*

99 *Id.* at 7.

100 *Id.* at 4, 7.

101 *Id.* at 5, 7 (noting that prediction accuracy is lower for districts with a small volume of data or with a large volume of data but irregular patterns due to local events).

102 *Id.*

103 *Id.* at 7.

better than that of others, and sometimes data from certain taxi companies—like Ghorbani and Zou's colon cancer patients from Nottingham—contributed strictly negative value such that it was better not to include that data at all.[104] Importantly, the variation in quality of contribution per taxi company was not closely tied to the volume of their data,[105] giving the lie to the use of data set volume as a measure of data set value.

It should be emphasized that, in both the cancer data and taxi data studies discussed above, the authors did not attribute variations in the data value to variations in data accuracy, another commonly cited metric in data marketplaces for assessing data value.[106] Rather, certain data relative to a certain task and conditioned on the presence of other data is simply more or less informative for a supervised machine learning model in understanding a pattern.[107] Or, in simpler words, certain data is more or less typical or representative of the reality that we are seeking to approximate with a particular model.

C. Summary of Observations

Not all data is equally valuable. The WTP and WTA valuations showcased above exhibit heterogeneity, often correlated with broad demographic categories. This supports the idea that variations in valuation for consumer data are common, regardless of the valuation mechanism or the definition of value. This is especially true for the consumer data that underpins the internet, given targeted ads' ubiquity as a source of website revenue.

Some data can even have negative value. As described *supra*, studies analyzing real-world data sets of taxi rides and colon cancer patients found

---

104 There is no substantial literature on the distribution of Shapley (or proxy) values for individual data points. However, the little data available in the literature is consistent with the results of this study on taxi company data. *See, e.g.*, GHORBANI & ZOU, *supra* note 87, at 6–7 (finding that adding individual data points with high estimated Shapley values can increase model performance, whereas adding data points with low estimated Shapley values can hurt performance); Ruoxi Jia et al., *Towards Efficient Data Valuation Based on the Shapley Value*, 89 PROC. OF THE 22ND INT'L CONF. ON A.I. & STAT. 1167, 1174 (2019), https://proceedings.mlr.press/v89/jia19a.html [https://perma.cc/T9BY-4YXA] (describing an experiment in which adding noisy data points decreased Shapley value of data).

105 GHORBANI & ZOU, *supra* note 87, at 5.

106 *See supra* notes 86–97 and accompanying text (describing studies assessing the value of data sets from different geographic or company sources by their contribution to performance of models that aggregate data from all geographies or companies).

107 Indeed, there is not yet a clear theoretical basis for understanding why the Shapley value is so variable, why high Shapley value correlates so strongly with model performance, and why certain data points would prove so much more valuable than others. For more background on data valuation and the Shapley metric, see generally Simons Institute, *What Is Your Data Worth? Equitable Data Valuation in Machine Learning*, YOUTUBE (July 8, 2019), https://www.youtube.com/watch?v=79pRqMq_-LE [https://perma.cc/GGA4-QSLR].

significant variation in the contributions of different data sets to model performance. The researchers found that the inclusion of some data sets could negatively affect model performance, and further that a data set's value did not strongly correlate with typical heuristics like volume of the data. *These findings suggest that the variation in consumer marketing WTP and WTA measures likely represents only a floor as to the magnitude of the variation in data value that can be measured with computational methods.* In other words, subjective and individualistic human valuations of data are, surprisingly, less variable than quantitative metrics of data value.

The relative contribution of value by data subjects will vary by *orders of magnitude* regardless of the form of valuation that is used. Such variation is found across a range of data sets and data-driven applications. As noted previously, men cost less than women in WTP terms, and Democratic donors cost less than Republican donors in WTA terms. Likewise, computational methods also identify huge variations in value in terms of contribution to model performance—variations that can *even highlight negative values of some data.*

Returning to the Reddit hypothetical introduced at the start of the work, data subject heterogeneity sheds light on just how varied fair compensation for redditors would be if they sought individualized payment for their data contributions. It could very well turn out that some redditors would receive ten or a thousand times the compensation of others. It could also turn out that some redditors *pollute* the data, in the sense of contributing negative value, not through ill will or low-quality data but simply because their contribution is unhelpful for a given task.[108]

II
DATA PROCESSOR HETEROGENEITY

Returning again to the motivating example of the relative value contribution of platforms and their user bases, as in the case of Facebook or Reddit, we next contemplate some indicia of the relative value contribution by a data controller or data processor—that is, by a firm or other entity that accesses data and uses it to achieve something. Most of the discussion here is focused—due to the necessary constraints of public data availability—on skill in developing algorithmic models. However, these insights likely also apply more generally, such as in skill exercised in the service of product design or community building, alongside any other key elements that can make or break an online community or digital firm.

---

108 This offers one manifestation of the phenomena described in Omri Ben-Shahar's theory of "data pollution." *See* Omri Ben-Shahar, *Data Pollution*, 11 J. LEGAL ANALYSIS 104, 105–06 (2019) (coining the concept of "data pollution" to describe the harm to social institutions and public interests caused by release and potential misuse of personal data).

In the previous section, we explored how model performance can reveal variations in the contributions of data sources and subjects. Model performance, however, varies not just with the quality of individual data points or subsets of data but also due to the skill of the practitioner and the degree of information provided by a data set in its totality. Let's now consider two elementary ways of observing variation that goes beyond the relative contributions of different data points.

First, we look to the performance of data processors conditional on access to the same data resources. We find substantial variation even in a highly compensated, performance-based task. This shows the important role of data processor skill for outcomes of algorithmic performance. Second, we look at the relative contribution of information in a data set versus skill of a practitioner across a range of tasks, as assessed by the variability in overall model performance, finding that the first order estimate of the relative contributions of the data set and the data processor can be quite variable. In other words, we likewise find variation in the relative value contributed by data subjects versus data processors in different tasks, analogous to heterogeneity in data value and in data processor skill. This discussion complements the one above, in showing that not all data processors or data sets are created equal.

A. Intertechnician Variation

Data controllers and data processors are not all created equal. On the one hand, this surely comes as an obvious statement to the reader. In prior boom years for Big Tech, it was well known that large firms paid substantial salaries for talent in AI, presumably reflecting a high premium for the most talented AI developers.[109] On the other hand, some data protection proposals, such as those that mandate easy and complete data exports or even data interoperability between platforms, seem to assume there is no innovation worthy of intellectual property protection in these data troves, or more specifically, seem to assume that the format of the data and the selection of what is recorded have little protected value. Let's briefly discuss both the topic of collecting and storing data, and separately the topic of analyzing data.

1. Skill in the Recording of Data

The choice of which information about the world, or about an abstract

109 *See* Cade Metz, *Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent*, N.Y. TIMES (Oct. 22, 2017), https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html [https://perma.cc/4KAR-FNFV] (discussing how technology firms are offering premium salaries to prospective employees with an AI background).

notion, to record is likely one of hard-fought experience, ingenuity, and some luck. Such a choice of data representation can be instrumental to a given task's level of success.[110] A topic that can only be given a hand-wavey description is the specific contribution data controllers make both in selecting what to record and selecting the format in which they record it. Yet this representation and selection of data likely represents a substantial contribution to the value of a given data set.

In my own experience, when seeking data from firms, I have found firms to be as protective of their data *formats* as they are of their data *contents*. Even when they have agreed to share information, they have requested that I generate an alternative *representation* of that data. Unfortunately, I have not been able to identify a public discussion of this to date, despite participating in numerous proprietary discussions of the problem.[111]

Some indicia of how data controllers contribute value by making choices of how to record data are depicted, nonetheless, in work presented by economist Laura Veldkamp, who recognized a progression of increasingly valuable data products, proceeding from "raw data" to "structured data" to "knowledge."[112] As Veldkamp depicts, in a figure reproduced in Figure 2, applying skilled labor (from a data manager or analyst, say) to a data record can drive that record into a higher value tier. Successive interactions with skilled laborers can draw "knowledge," as Veldkamp calls it, from "raw data."[113]

110 *See generally* Edward L. Fink, *The FAQs on Data Transformation*, 76 COMMC'N MONOGRAPHS 379 (2009).

111 The author would appreciate any recommendations in this area. Speculatively, it may be so difficult to find public discussions of data representations because they are so heavily protected by firms as valuable forms of intellectual property.

112 Financial Management Association International, *Valuing Data as an Asset - Laura Veldkamp*, YOUTUBE, at 27:00 (Feb. 22, 2023), https://www.youtube.com/watch?v=l_oDLKha4co [https://perma.cc/96RM-PV2X] (recording a presentation by Laura Veldkamp).

113 *See id.* at 28:30 (describing how firms add more value to data via this process).

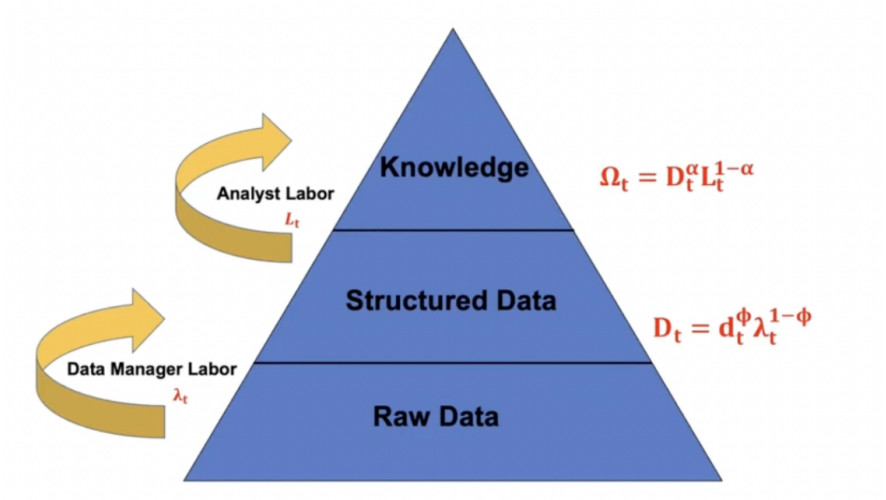

Figure 2: A proposed flow of ever-more improved and ever-more valuable forms of data by Veldkamp (2022).

However, I would propose that Veldkamp's model is incomplete. First, as many have before, I posit that the notion of "raw data" is troubling, and obscures the many decisions and conceptual underpinnings that necessarily inform any digital representation that purports to have translated information about the world into a digital format.[114] The choice of how to represent information—that so called "raw data"—entails many choices: which information to record and how to organize that information is a significant part of how data controllers exercise skill in the collection of so-called raw

---

114 *See, e.g.*, Nick Barrowman, *Why Data Is Never Raw*, NEW ATLANTIS, Summer/Fall 2018, at 129, 133–34, https://www.thenewatlantis.com/publications/why-data-is-never-raw [https://perma.cc/9DNU-UJMC] (describing the problems with the term "raw data"). Indeed, the emphasis that continues to be placed on describing what are in fact substantial investments of ingenuity and skill as raw resouces continues to be a theme in policy discussions about tech. *See, e.g.*, Lina Khan, *We Must Regulate A.I. Here's How*, N.Y. TIMES (May 3, 2023), https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html [https://perma.cc/5MKB-7C7M] (referencing "necessary raw materials" with a URL link that led to an announcement by Google of a new supercomputer, hardly a raw material). Chair Khan's statement was troubling not only for a seemingly polemical invocation of the "raw" quality of these resources but also for the conclusory statement that such computing resources are necessary, a statement that is far from clear as a description of empirical reality. *See, e.g.*, Rohan Taori et al., *Alpaca: A Strong, Replicable Instruction-Following Model*, STAN. CTR. FOR RSCH. ON FOUND. MODELS (Mar. 13, 2023), https://crfm.stanford.edu/2023/03/13/alpaca.html [https://perma.cc/UUM8-FJL3] (presenting a model much smaller than OpenAI's LLMs but, in early tests, offering a similar level of performance). Although Chair Khan may prove to be right in assuming the necessity of access to certain forms of IP for success in cutting edge technology, such materials are not raw, and Chair Khan—like everyone—can only offer speculative claims for the moment.

data.

To return to our thought experiment as to Reddit and the redditors, one can imagine several ways in which Reddit has exercised analogous skill. In addition to the data engineering effort necessary to store the complex nested structure of ever more tangential conversations all within the same discussion thread, Reddit has also created forms of metadata both about the discussant and about a particular comment in the form of various badges and a system of upvoting and downvoting, as well as the propagation of a culture in which upvotes and downvotes are themselves regulated, likely to increase the informative value.[115] Further, we have no way of knowing what additional data, if any, Reddit has recorded and in what format apart from what we can see manifest in the user interface. It is possible that the company records additional information or makes unseen inferences that can guide its presentation of content.[116] In short, even "raw data" represents a substantial exercise of skill, ingenuity, and hard-won experience, all of which are commercially valuable and (often) legally protectable.

2. Skill in the Modeling of Data

It has not been easy to identify publicly available examples of variation in the skill of selecting and recording data. On the other hand, there are readily available public forums to learn about skill in the modeling of data, and some of the data from such public forums has been analyzed and presented in Figure 3. This figure was generated with the leaderboard contents from a Kaggle competition recently hosted to automatically label frames of a video in which NFL players had come into contact.[117] The platform has also branched out to include educational materials and other forms of content.

One hallmark of Kaggle competitions that makes them particularly interesting is that sometimes there is serious skin in the game. For example, in the data analyzed in Figure 3, there was serious prize money at stake, with $100,000 of prizes total, of which $50,000 was for first prize alone. Nearly 1,000 teams competed over the course of three months for the prize money.

---

115 For example, some subreddits limit how many times a user can upvote or downvote in a given time period, injecting scarcity into the dynamics of voting.

116 For example, one can imagine that Reddit might record behavioral data such as speed of content inputs, to identify and weed out non-human participants on the platform.

117 Kaggle is a large online platform that offers both training in machine learning and access to various data sets, as well as competitions. *See, e.g.*, *1st and Future – Player Contact Detection Leaderboard*, KAGGLE (Mar. 1, 2023), https://www.kaggle.com/competitions/nfl-player-contact-detection/leaderboard [https://perma.cc/FT7A-D8BD] (describing a competition hosted on Kaggle and the accompanying data set). It is a widely known platform, with millions of registered users. *Kaggle Has 10 Million Registered Users!*, KAGGLE (June 18, 2022), https://www.kaggle.com/discussions/general/332147 [https://perma.cc/H88E-CNKE].

There is every reason to think that the teams took this competition seriously both given the money at stake and also given the reputation Kaggle has established as a place for machine learning engineers to prove their skill level.[118]

It could be that there is relatively little skill involved in generating ML models and that most of the value lies in the data itself. If there were little skill involved in crafting ML solutions, we would expect to see a fairly narrow set of results because participants all have access to the same data.[119] But this is not what we see.

Rather we can easily identify evidence strongly suggestive of the important role of ML operator skill in crafting good models. As shown in Figure 3, however, there is a substantial range in the outcome that was used to evaluate the winners. The X-axis depicts a particular value of correlation, where the correlation is measured between correct outputs and a model's outputs. For each Kaggle participant in Contest 1, the correlation for their model and the correct outputs was calculated. Then a histogram of all correlations for all participants—counting the correlation of each participant once—was constructed. This histogram thus gives a view of the distribution of model performances across all Kaggle participants in Contest 1.[120]

Reading this histogram, we can see that the range of model performance varied between approximately 0.55 to 0.8 correlation, with possible values between -1 (reliably always wrong) and 1 (reliably always correct). As can be seen in the histogram, there is a prominent modal value—that is, the most common value—at around a correlation of 0.7. A reasonable inference seems to be that this performance value represents the outcome of modeling with the typical level of skill and effort devoted in that contest by Kaggle participants. This performance rate is, however, more than 10% worse than the top performers who reached correlation levels of 0.8, showing that there is indeed a role for skill and effort in developing ML models.

---

118 Many prominent machine learning practitioners have come up through the ranks of Kaggle. Consider Jeremy Howard, now a founder of multiple AI ventures, who came to prominence as a top competitor on Kaggle. *See Jeremy Howard*, KAGGLE, https://www.kaggle.com/jhoward [https://perma.cc/M8AE-DHXA] (last visited Nov. 21, 2023) (outlining Jeremy Howard's other business ventures and displaying his success on Kaggle leaderboards).

119 *See 1st and Future – Player Contact Detection Leaderboard*, KAGGLE, https://www.kaggle.com/competitions/nfl-player-contact-detection/leaderboard [https://perma.cc/5KM7-5VQ2] (last visited Nov. 21, 2023) (demonstrating the wide variety in scores in a competition where all participants have identical data).

120 In fact, the figures show the distribution of the top 90% by performance of participants. The lowest 10% of performers were cut from the representation to account for the possibility that some teams had dropped out or otherwise possibly stopped giving a good effort before the end of the competition.

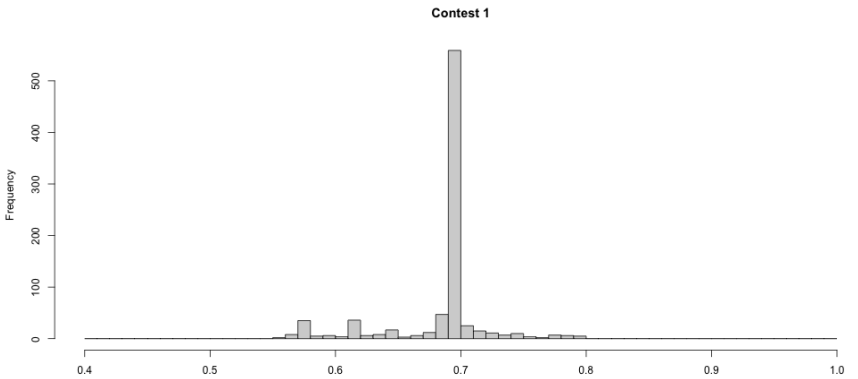

Figure 3: A spread of model performance metrics in one highly incentivized Kaggle competition.

The wide range of skill evidenced by the differential model performance rates shown in Figure 3 provides convincing evidence that there may be some modal level of skill that correlates with the modal level of model performance. But it also provides convincing evidence that there are outliers above and below that mode. While skeptical readers might discount the left tail of the distribution as participants who dropped out or were disinterested, the right tail suggests a meaningful spread in skill, even in a highly incentivized environment.

In larger scale empirical studies, gaps in the ability of a data controller or processor to use data effectively to accomplish a task have also been recognized, and even quantified. In estimating the value of data in the hands of different firms, Laura Veldkamp has estimated that—where firms have different data resources and tech talent—the value to those firms of the data set could range between ten dollars and over one million dollars for the same data, a variation of five orders of magnitude.[121] Such an empirical data point suggests that the variation seen in Figure 3 is only the tip of the iceberg regarding the relative value contribution of different data processors for the same task.

Previous scholarship has not wholly negated that there are differentials in skill. Nonetheless, little scholarship has taken seriously that there are skill differentials and that these might even be tied directly to the benefits realized. Once we see these differentials and contemplate them seriously, it

121 *See* Maryam Farboodi, Dhruv Singal, Laura Veldkamp & Venky Venkateswaran, *Valuing Financial Data*, 29, 35–36 (Nat'l Bureau of Econ. Rsch., Working Paper 29894, 2022), https://www.nber.org/papers/w29894 [https://perma.cc/82V7-MM69] (modelling values of data sets to theoretical firms with different characteristics).

should prompt us to think more carefully about what these differentials mean for privacy, competition, and most particularly for innovation. If it turns out that the forms of privacy regulation we have chosen are likely to dampen incentives for innovation, and if we believe that there is true skill and ingenuity exercised in most applications that drive the Big Data economy, we must better understand the relationship between our current and proposed privacy regulations and potential implications in a market comprised of heterogeneously skilled data processors. Certainly, we'll want to find ways such that, when sensitive data is used in appropriate circumstances, it is also likely to be allocated to more skilled data processors, all else equal.

To bring this back to the thought experiment from the introduction, the mapping is of course quite obvious and direct in this case. Where Reddit seeks to position itself not as merely licensing the content of redditors but in fact contributing something of value through its recording and elicitation of the data—and likewise where site moderators who don't produce content might likewise seek credit for the high value of the data, rather than for producing data themselves—data such as that discussed here is likely to buttress their case that they have exercised skill and added value to the modeling task. Reddit hasn't been a dumb and mindless operator of a data collection site. It has been a skilled operator—presumably one of the best ones. Any site could record content, but arguably only Reddit and its hardworking, volunteer moderators have exercised the right degree of skill, creativity, and ingenuity so as to produce and record such high value data—data that is far from raw but that has in fact been engineered into its putatively high value. In short, data from a real world highly incentivized modeling task empirically backs up the notion that much skill is contributed by data controllers and data processors, apart from whatever "raw" value might be in the data itself.

B. Some Inklings on the Relative Contributions of Data and Data Processor

Quite a lot of skill comes not in directly training models but rather in preparing the inputs for those models. Such preparation runs all the way back to the "capture" of "raw data." This is a factor that has come to be increasingly understood by the AI community, with a newly named movement of "data-centric AI," which in its simplest form posits that improving quality of data can be instrumental in improving AI performance, replacing an earlier emphasis in the AI community on *quantity* of data and *complexity* of architecture.[122] In other words, something about the essence of

122 *See* Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang

the data (whatever that means) contributes to how useful it can be, and this will vary for different data sets and different tasks.

Going again to the public data available on Kaggle, we can easily find that this is borne out in the data. Consider a very simple set of examples assembled by considering recent Kaggle competitions in which at least $100,000 in prize money was available.[123] In such contests, the participants are likely motivated to do their utmost with the data available. Further, participants all have access to the same data set. It seems likely that with highly motivated and well qualified participants and in a competitive setting, that something close to a state-of-the-art optimum level of performance will be achieved in such competitions. And yet, conditional on a given data set, what that performance is will vary wildly. In the following plot, five recent competitions were considered, in each case with a scoring metric that ran between -1 and 1 as potential values.

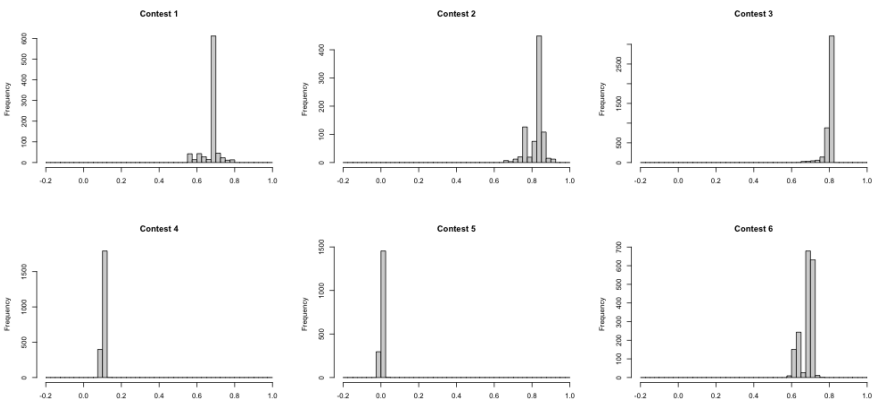

Figure 4: A spread of model performance metrics in six highly incentivized Kaggle competitions.[124]

& James Zou, *Advances, Challenges and Opportunities in Creating Data for Trustworthy AI*, 4 NATURE MACH. INTEL. 669, 669, 674 (2022) (discussing the benefits of data-centric AI research and how it is under-researched).

123 The prize money and recency were the only selection criteria. No selection of the competitions was imposed based on the spread on their leaderboards.

124 These data represent the reported final performance values in six Kaggle competitions, each of which was completed recently (as of the writing in May 2023) and each of which featured prize funds of at least $100,000. *1st and Future – Player Contact Detection Leaderboard*, KAGGLE, https://www.kaggle.com/competitions/nfl-player-contact-detection/leaderboard [https://perma.cc/5KM7-5VQ2] (last visited Nov. 21, 2023) (Contest 1); *Google AI4Code – Understand Code in Python Notebooks Leaderboard*, KAGGLE, https://www.kaggle.com/competitions/AI4Code/leaderboard [https://perma.cc/XBW5-LML2]

The number of participants in the competition ranged between 940 (Contest 1) and 4,876 (Contest 3) participants, in each case a substantial sampling of modeling participants.[125] The spread in the distribution is not down to simply having more participants and so seeing a wider range of skill, as exploratory data analysis revealed that the spread in performance did not correlate with number of participants. Rather this spread tells us something about how much the degree of skill mattered in a particular application.[126] Skill—that is the range of outcomes by data processors conditional on access to the same data set—appeared to matter far more (result in more variation) in some modeling tasks as compared to others. Consider that sometimes the winning performance level was substantially above the modal performance bin, as in Contests 1, 2, and 3, but that in some cases, as in Contests 4 and 5, the modal bin also contained the highest performance level, suggesting—assuming the bins represent a reasonable degree of granularity[127]—that in some cases that data itself is indeed doing all the work and data processor skill matters little.

Of course, some of the appearance of the distribution will come down to the size of bins chosen from the histogram, but that is not what is driving the overall observation of quite different distributions for different tasks.

(last visited Nov. 21, 2023) (Contest 2); *American Express – Default Prediction Leaderboard*, KAGGLE, https://www.kaggle.com/competitions/amex-default-prediction/leaderboard [https://perma.cc/BH5B-S5BD] (last visited Nov. 21, 2023) (Contest 3); *Ubiquant Market Prediction Leaderboard*, KAGGLE, https://www.kaggle.com/competitions/ubiquant-market-prediction/leaderboard [https://perma.cc/2XZD-SWKW] (last visited Nov. 21, 2023) (Contest 4); *G Research Crypto Forecasting Leaderboard*, KAGGLE, https://www.kaggle.com/competitions/g-research-crypto-forecasting/leaderboard [https://perma.cc/JQ5N-J97E] (last visited Nov. 21, 2023) (Contest 5); *Feedback Prize – Evaluating Student Writing Leaderboard*, KAGGLE, https://www.kaggle.com/competitions/feedback-prize-2021/leaderboard [https://perma.cc/5XN5-K8U2] (last visited Nov. 21, 2023) (Contest 6).

125 *Supra* note 124 and accompanying text.

126 The spread in the distribution also tells us how difficult a particular problem was and how far from perfect the ultimate models performed. One way to get at how much "skill" participants added at all relative to the data itself in its "raw" form (that is, the captured form in which it is provided) would be to find a representation of what is usually known as a null model. That is, a model that is dumb and not responsive to features but in some way incorporates the bare minimum of information provided by the data. For example, one version of a null model (posited as a model to beat) would be for time series predictions to, for each time period's forecast, use the most recent value and carry it forward. Or for categorical data, an example of a null model would be to always predict the most common class. The performance of such null models provides a way to establish the bare minimum performance one would expect doing no more than some kind of informed guesstimate based on aggregate information. Thus, the null model establishes the floor of model performance and perfection establishes the theoretical, though usually not achievable, ceiling.

127 Whether this is the case will depend on factors of how impact relates to model performance. For example, if the model is being used in a way that affects fundamental human rights—perhaps as in an application in the criminal justice system—there could be normative concerns that would emphasize every last differential in performance. On the other hand, for many real-world applications, "good enough" will dictate little changes for the very same range in performance metrics.

Rather, the key factor is that the distribution of performance varies in a way that makes some suggestion both as to the importance of skill of the data processor and shows that this is likely to be highly varied, even in a highly incentivized and highly competitive environment like a $100,000 Kaggle competition. In short, how much a data processor can be on "auto pilot" rather than exercising true effort and skill, will partly depend on the value of the data for a task. In some cases, the data will offer a slam dunk, while in other cases there may be hidden opportunities, but only for a good data processor. Ultimately, there is valuation heterogeneity at every step in the modeling task.

III
IMPLICATIONS FOR LAW AND POLICY

As with widgets or employees, it turns out that data subjects and data processors are not all created equal. Yet scholars focused on privacy, competition, or innovation appear agnostic as to the reality of heterogeneous data values or their import. If we re-examine basic elements of current law and policy discussions about privacy, competition, and innovation, we see a lack of appreciation for the value that may be gained or lost in recognizing value heterogeneity as among data subjects and data controllers. Likewise, we observe that some recent changes or recommended policy proposals may carry hidden costs related to data value heterogeneity.

A. Privacy

Despite a decade of anxiety about the Big Data economy, American law remains permissive. Time and again Americans have indicated fear and a sense of powerlessness and ignorance as to the degree of information collected about them by the private sector. For example, a 2019 Pew Research survey found that most Americans believe that the potential risks of data collection outweigh the benefits received and that they do not benefit from data collection.[128] For the most part, consumer protection laws allow data controllers to set their own privacy standards so long as those standards are transparent and are honored, somewhat like (but distinct from) contract terms. This regime is typically described as notice-and-consent, a regime of privacy self-management that has long been criticized as inadequate for purposes of informing consumers or providing meaningful privacy

128 Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar & Erica Turner, *Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information*, PEW RSCH. CTR. (Nov. 15, 2019), https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information [https://perma.cc/LXF5-39DJ].

choices.[129]

Data heterogeneity provides yet another reason to justify the intuitions of ordinary Americans as to the inadequacy of the benefits they receive in exchange for their data. The orthodox view of exchanging privacy for services and innovation is that giving up some degree of privacy makes the world a better place: Data subjects enjoy convenience, and firms use collected data to innovate to society's benefit. One way to understand Americans' frustration with current practices in the Big Data economy is that they believe they are not getting a fair deal in this exchange. Data subject heterogeneity supports this intuition. Through no fault of their own, many—maybe even most—Americans' data *isn't even useful to refining a product or creating new algorithmic intelligence*. The collection of many Americans' data may add little or nothing to product or service improvement; their data may very well be collected in vain.

Contemporary commercial data collection may be driven by speculative notions that data may be useful at some indefinite point in the future.[130] Thanks to a permissive U.S. privacy law regime, the cost of collecting and retaining data in the meantime is low.[131] Given that the volume of data produced and recorded in ordinary consumer contexts continues to expand rapidly[132] despite consumer unhappiness with these practices (and a pervasive threat of data breaches and resulting risks of identity theft), it is incumbent on firms and policymakers alike to determine whether there are any benefits at all to storing much of that data. Imposing *data minimization* requirements on firms would push firms to make such a determination. A principled means of valuing data—and there are many available, as discussed *supra*—could serve as a criterion for firms to prioritize what data to remove or retain. That is, stringent requirements would force firms to actively assess which data is useful for their task, and to retain only data valuable for a commercial purpose, rather than all otherwise legally-available data.

---

129 *See generally* Daniel J. Solove, *Murky Consent: An Approach to the Fictions of Consent in Privacy Law*, 104 B.U. L. REV. (forthcoming) (manuscript at 11–33) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4333743) (providing an overview of the current consent fiction and proposing changes).

130 Based on discussions with people in the industry, even the largest and most sophisticated firms mostly collect data that is not useful for their models, but is instead collected for future speculative possibilities. Most of these admissions are made at closed-door meetings, making it difficult to cite a public source for this proposition.

131 *See* Yafit Lev-Aretz & Katherine J. Strandburg, *Privacy Regulation and Innovation Policy*, YALE J. L. & TECH. 256, 294 (2020) (stating that the benefits of data collection almost always outweigh the costs because of relaxed data privacy laws).

132 *See* Petroc Taylor, *Data Growth Worldwide 2010-2025*, STATISTA (Nov. 16, 2023), https://www.statista.com/statistics/871513/worldwide-data-created [https://perma.cc/JUF4-QXGJ] (discussing the rate of data growth and providing projections).

An interesting and underappreciated fact about the heterogeneous value of data subjects is the fact that for many purposes, such as training AI models, current industry practice of collecting data from all individuals and buying, selling, and trading whole data sets is probably inefficient with regards to training good models. Those seeking to reduce data flows[133]—one manifestation of privacy by design—would find some purchase in citing the highly unequal value of different data points. Privacy advocates could ask, on a quantitative basis as well as on a normative basis, whether quite so much data needs to be collected and retained. From both a dignitiarian and a utilitarian perspective, data controllers could and should be collecting and modeling far less data than they currently do.[134] If privacy law's duties of data minimization reflected an understanding of data heterogeneity, this could incentivize data processors to collect and hold far less data and lead to a truer form of data minimization.[135] In other words, some implications of data value heterogeneity point to the desirability of more stringent and pro-consumer privacy protections. A world where data processors pay, and data subjects receive, fair value for data would likely involve decisions by data processors not to obtain data from uninformative subjects. It would also likely involve processors obtaining valuable data at its fair cost, within the limits of consent or another normative framework, rather than the low, flat pricing structure we typically see in industry right now.[136]

Not all considerations of heterogeneity in data value creation point in a pro-consumer direction. If we take seriously the skill of data controllers in creating opportunities to generate data and representing that data, there are

133 I recognize that not all forms of imaginable privacy regulation need tend towards the negative of ever-reduced data flows. For example, Helen Nissenbaum's theory of contextual integrity provides for *appropriate* data flows, which might sometimes even entail more rather than less data flow relative to current practice. *See* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE, 129–30 (2010) (discussing contextual integrity and the ideal amount of data flows with respect to privacy). However, the comment here is specifically with respect to certain practices or ideals pursued by some in cybersecurity or in privacy regulation.

134 *See* Lev-Aretz & Strandburg, *supra* note 131, at 291 (discussing how dignity values should be preserved outside of any cost-benefit data analysis).

135 The General Data Protection Regulation has incorporated provisions that do promote practices that would lead to less collection and retention of data through the principles of data minimization and purpose limitation. *See* Council Regulation 2016/679, art. 5, 2016 O.J. (L119) 1(data minimization); Council Regulation 2016/679, art. 25, 2016 O.J. (L119) 1 (purpose limitation). However, those principles do not explicitly contemplate the additional empirical wrinkle presented here, that some data is more or less value-producing. This empirical fact can further strengthen expectations of data minimization and purpose limitation. Surely, those data points shown to have a negative Shapley value might compel even more drastic steps towards data minimization than data with a positive Shapley value where data collection has been specified for a particular purpose. So the computation of Shapley value, and other data valuation techniques, give more content and more quantifiability to these existing requirements.

136 See *supra* Section I.A(2) discussion on flat or volume pricing by data brokers for any given data set.

also reasons to question the basis for data access laws that require firms to make all data pertaining to a data subject available for easy export. Data export requirements have appeared in new wave privacy legislation, including GDPR's accommodation of subject access requests and recent state laws.[137] Some data structures—the representation of data in logical terms—may be patentable.[138] Likewise, methods for displaying information are patentable.[139] Even when not patentable, they surely benefit in some cases from trade secret protection. This is because decisions about what data to record and how to record that data do indeed represent true, value-creating skill. To the extent data controllers do currently fully comply with data access and export rules, these rules may be disincentivizing them to put effort into improving their data—even as the machine learning community increasingly appreciates the importance of high-quality data.[140]

Taking seriously the intellectual contributions that data controllers make to their data raises two questions. First, is it either a fair deal or at least economically efficient to require data controllers to share all data they have about data subjects? Perhaps, instead of blanket access and export requirements, regulators should establish pre-defined categories and the formats for those categories to be shared. This would both assist with interoperability (discussed *infra*) and also protect intellectual investments, possibly a better balancing of privacy and innovation than a blanket export requirement.

Finally, this work should not be misconstrued or taken to conflate the value of data with the value of or merit of privacy rights. This work observes that, in a given context and in the presence of a given set of data and a given task, the data of different data subjects will contribute differentially to model performance, and that likewise different data processors will show different levels of skill in fulfilling a task. *None of this means that certain data subjects are more or less deserving of privacy*. Nothing in this work should be taken to imply a regime, for example, of privacy eminent domain whereby data subjects whose data is more informative should be forced to waive their rights to privacy in a way that less informative data subjects are not, or to otherwise have their fundamental rights infringed simply because their data

137 *See* Council Regulation 2016/679, art. 20, 2016 O.J. (L119) 1 (outlining the rights of data subjects to data portability); *see also* Andrew Folks, *US State Privacy Legislation Tracker*, IAPP, https://iapp.org/resources/article/us-state-privacy-legislation-tracker [https://perma.cc/5Z48-ZLWP] (last updated Nov. 10, 2023) (mapping privacy laws across the United States).

138 *See* Andrew Joseph Hollander, *Patenting Computer Data Structures: The Ghost, the Machine and the Federal Circuit*, 2 DUKE L. & TECH. REV. 1, 6–11 (2003) (providing an overview of how federal courts have addressed the patentability of data structures).

139 *See, e.g.*, *Hyperbolic Tree Space Display of Comput. Sys. Monitoring and Analysis Data*, U.S. Patent No.7,143,392 (filed Feb. 11, 2002) (providing an example of a patented data structure).

140 *See* Liang et al., *supra* note 122 (describing the benefits of focusing on the quality of data).

is judged most valuable.[141]

A limitation on the conclusions to be drawn from this work is that this work presupposes a normative framework that accepts the exercise of *valuing* data or inferences with reference to the economic premiums placed on the data or inferences. At the end of the day, it remains a very fair question to ask whether the *value* of data lies really in its utility or whether such questions should be fully dependent only on answering more fundamental dignitarian questions as to the normative acceptability of such exercises.[142] For those who have notions of privacy that do not admit the legitimacy of balancing privacy against other values in the legal system, it will necessarily be of little interest or value what the relative value contributions are.

B. Competition

Data value heterogeneity is relevant to competition policy in very similar ways to the concerns about privacy. First, it suggests that concerns about "data moats"—that actors will be unable to enter new markets because they lack data—may be overblown. In some cases, not much data is needed to accomplish significant tasks, and large chunks of data may even be harmful to accomplishing a task. Firms that succumb to data fetishism rather than smart use of data may face a competitive disadvantage. It may very well be that so long as potential new entrants can access some data, they will be reasonably competitive even as new market entrants. The influence of data

---

141 In some ways this debate is comparable to that which ensued after rising awareness regarding the particular contribution of cells taken from Henrietta Lacks's body and, without her knowledge or consent, used to propagate a highly successful cell line for research. *See Henrietta Lacks: Science Must Right a Historical Wrong*, 585 NATURE 7, 7 (2020) (providing policy recommendations in light of Henrietta Lacks's story). Likewise, *Moore v. The Regents of University of California*, 793 P.2d 479, 480, 482 (Cal. 1990), presents a canonical treatment of the non-consensual use of tissue from an individual who turned out to be extraordinary in potential contribution to medical science. To date, in the biomedical community such moral reckonings have largely focused on consent rather than compensation, in part due to fears that compensation could become so high as to undermine consent. *See generally* Julian Savulescu, *The Fiction of "Undue Inducement": Why Researchers Should Be Allowed to Pay Participants* Any *Amount of Money for* Any *Reasonable Research Project*, 1 AM. J. BIOETHICS 85 (2001).

142 *See* Tal Z. Zarsky, *The Privacy Innovation Conundrum*, 19 LEWIS & CLARK L. REV. 115, 143–44 (2015) ("The online pornography and the gaming (a.k.a. gambling) industries are both known technological trailblazers. . . . So in view of all this, would it be acceptable to argue that regulations harshly limiting the legality of and accessibility to these technological industries should be softened so as to promote online innovation? One would be hard pressed to find any policymaker seriously advocating this position. The negative aspects of both pornography consumption and gambling activity—even when carried out legally by autonomous adults—will intuitively trump any of the benefits resulting from the innovative practices noted."). It is worth noting that not everyone would necessarily accept Zarsky's logic that it is obvious that it is not worth promoting online gambling or prostitution services even if they are associated with innovation. Nor is it obvious that Zarsky is wrong, and therefore, for purposes of this article it is likewise not obvious on a normative basis that data valuation is of little or no interest as a proxy for innovation and counterbalance to privacy interests.

value heterogeneity has not been discussed previously in the competition law literature, but understanding that some data are more valuable than others reveals that data volume may not be as severe a barrier to entry as previously thought.

The observation of data value heterogeneity also speaks directly to questions posed by newly enacted regulations, particularly the Digital Markets Act (DMA) and Digital Services Act (DSA) from Europe. The Digital Markets Act (DMA) Article 5(2) prohibits various scenarios by which a gatekeeper might combine personal data from one core service to another.[143] For example, the Act would prohibit Meta from using data collected on Facebook to target ads on Instagram.[144] Recognizing data value heterogeneity permits two insights about the DMA. First, it may very well be that Facebook may not suffer significant losses in performance or profit from this new statutory limitation because big data is unnecessary for successful algorithm development.

Second, even if the regulation succeeds in achieving its policy goals, it may have done so for the wrong reason. Various EU regulations—the GDPR, DMA, DSA, and the Data Act, for example—attempt to make data interoperable and more readily available by imposing public access rights to some forms of data and by further requiring that the data take an easy-to-use standard machine-readable format.[145] These de facto and de jure interoperability requirements may be improperly targeted, especially where they focus on ensuring that *all data* be made available and especially where they emphasize *ease of access* to that data.

By focusing on large volumes of personal data, the regulation fails to distinguish among many functions of data, and particularly fails to distinguish between innovative uses of data and bare personalization. The latter rote uses of personal data adds little or minimal value and may be intrusive or abusive. It may very well be that any competitive advantage does not come from innovation returns to data but from bare personalization. In this case, legislators would do better to tailor acts according to data use rather than merely according to platform size. As this paper shows, size is often not a good proxy for value or power.

---

143 *See* Council Regulation 2022/1925, art. 5, 2022 O.J. (L265) 1 (outlining restrictions on what gatekeepers may do with data).

144 Colin Wall & Eugenia Lostri, *The European Union's Digital Markets Act: A Primer*, CTR. FOR STRATEGIC & INT'L STUD. (Feb. 8, 2022), https://www.csis.org/analysis/european-unions-digital-markets-act-primer [https://perma.cc/83T6-TPJ3].

145 *See, e.g.*, Martin Braun, Anne Vallery & Itsiq Benizri, *Details of the EU Data Act (1)—Data Access Rights and Obligations*, WILMERHALE (Dec. 4, 2023), https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20231204-details-of-the-eu-data-act-1-data-access-rights-and-obligations [https://perma.cc/HS3Y-M537] (summarizing the data access and interoperability requirements under the Data Act and how they interact with the GDPR).

The heterogeneity of data value also offers information relevant to the likely effects of the Digital Markets Act (DMA) and the Digital Services Act (DSA). Both laws impose data reporting and access requirements—effectively interoperability requirements—and both use language suggesting that all data should be reported where the requirements adhere.[146] Heterogeneity of the value of data combined with the surprising smallness of data actually needed to accomplish some tasks raise legitimate questions as to whether the substantial data provision requirements in the DMA and DSA—that all data (rather than samples of data) be made available to researchers under certain circumstances—will achieve their target. Although the regime is far too new for the concrete implementation details to be established, it seems likely that the very bigness of the data sets that researchers will inevitably request and as putatively required by the DMA and DSA statutory requirements could themselves be used pretextually[147] to heighten the barriers for regulators and academic researchers to access the information. For example, in the interests of keeping a reasoned request attainable in the DSA, firms might object to requests for all data or large sets of data and might therefore be able to escape requests in the name of protecting cybersecurity or other such privacy pretexts. If researchers appreciate this fact and find ways to specify requests for high quality rather than high volume data, a small subset of data could be nearly or equally as informative for some tasks.[148] This smaller request could in turn reduce the cybersecurity concerns related to transferring large troves of data (or making them accessible more widely) and could likewise lower the barrier to entry for the qualifications researchers would need to handle and draw meaningful conclusions from forthcoming platform data. The descriptive portrait offered in this work suggests that researchers and regulators alike would not typically need access to *all* platform data but rather to only a subset of

146 *See generally* MARC BOURREAU, DMA HORIZONTAL AND VERTICAL INTEROPERABILITY OBLIGATIONS (2022), https://cerre.eu/wp-content/uploads/2022/11/DMA_HorizontalandVerticalInteroperability.pdf [https://perma.cc/6QLM-XNBP] (outlining the DMA's rules on interoperability); Alex Engler, *Platform Data Access Is a Lynchpin of the EU's Digital Services Act*, BROOKINGS (Jan. 15, 2021), https://www.brookings.edu/articles/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act [https://perma.cc/9KXQ-6WCM] (noting how the DSA requires websites to provide researchers access to data).

147 *Cf.* Rory Van Loo, *Privacy Pretexts*, 108 CORNELL L. REV. 1, 38 (2022) (explaining how companies use protecting privacy as pretext for blocking government and researcher access to data in an American legal context).

148 Questions as to who would establish such criteria or apply them—especially given widespread mistrust of the firms associated with the large platforms to which Article 31 applies ("very large, online platforms") might render this attempt futile for reasons of a lack of trust or credible means of verification.

informative data.[149]

On the other hand, an appreciation of the heterogeneity of data may point to an even larger competition problem than has so far been discussed. Future empirical research should look in novel ways, premised on the strong variation in data subject contributions, at questions of whether more valuable data is ever rewarded with more valuable service. That is, to the extent that firms allege that consumers are happily trading data for services, can they also show that these trades track user heterogeneity in data value contribution? Do more valuable data subjects receive more valuable products, perhaps better performing algorithms or more bonuses for more intense usage of a service? And when data can be valued, are consumers receiving fair value for data? If not, this could point to a new way of understanding manifestations of monopoly or monopsony power in digital markets.

C. Innovation

Variation in the contributions of data subjects and data controllers also informs possibilities for innovation law and policy. One can contemplate how compliance with data-related law may reflect innovation concerns. Data value heterogeneity adds a new dimension to the latter question: Where firms are deviating from privacy law requirements, might this be in the interest of protecting innovation? Consider data access, a right increasingly mandated around the world, for example in the European Union's General Data Protection Regulation[150] as well as the California Consumer Privacy Act.[151] Recent work suggests that when law requires data controllers or processors to disclose data in response to requests, they generally release less data than they have collected. That is, they under-release. In contrast, firms appear to adhere more closely to other elements of data law that do not require them to share their data contents and format. For example, two empirical studies of compliance with GDPR requirements found high rates of compliance with

149 Of course, critics could point out that, if firms are left to determine which subset of data to share, they would have incentives to manipulate the selection of that data. Nonetheless, in some ways it seems technically more approachable to regulate appropriate selection methodologies for informative subsets of data rather than to deal with the cybersecurity challenges of making *all* data for a particular modeling problem available to regulators or researchers. Of course, if some tasks are seeking to establish a comprehensive census, there is potentially a legitimate need for all data. On the other hand, one can imagine many valid and socially important research questions that do not necessitate counting exhaustively every instance of a particular phenomenon of interest.

150 Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 15, 2016 O.J. (L 119) 1.

151 Cal. Civ. Code § 1798.115 (West 2023).

data deletion commands—73% in one study[152]—but lower rates of compliance—only 29% in another study[153]—when it came to data export under the GDPR.[154] Similarly low rates of compliance with data export rights under the California Consumer Privacy Act—with less than 20% compliance[155]—suggest that the trend towards especially low compliance for data export rights could be an international trend. These data do not prove the hypothesis that privacy law compliance partly reflects innovation concerns, but the data certainly are consistent with such a hypothesis and show the need for further empirical study.

This empirical work suggests that it may very well be that firms' innovation regarding improving the value of data by refining data storage formats and targeting particular forms of information for recording that is more valuable than any particular piece of information itself. In other words, the know how regarding how to work with data and what data is informative may be far more valuable than the data itself. As assessed here in an emerging body of compliance studies, it seems that firms may be more reluctant to export data and make it available in their chosen format than they are to delete that data altogether. This is consistent with an interpretation that firms are holding data that is not especially valuable to them (but that they fear may be valuable to others). It is also consistent with an interpretation in which the firm views its IP in the selection of data recorded or format of that data as significantly valuable even apart from the data itself.

Another fascinating empirical finding comes from a 2021 study by Emmanuel Syrmoudis and colleagues in which the authors studied the

---

152 Eduard Rupp, Emmanuel Syrmoudis & Jens Grossklags, *Leave No Data Behind – Empirical Insights into Data Erasure from Online Services*, 2022 PROC. ON PRIV. ENHANCING TECHS. no. 3, 2022, at 446.

153 Sophie Kuebler-Wachendorff, Robert Luzsa, Johann Kranz, Stefan Mager, Emmanuel Syrmoudis, Susanne Mayr & Jens Grossklags, *The Right to Data Portability: Conception, Status Quo, and Future Directions*, 44 INFORMATIK SPEKTRUM 264, 268 (2021) (concluding that only 28.6% of service providers in the study sample were compliant with GDPR Art. 20 (1)).

154 There are some reasons why the two numbers are not directly comparable, although there was an author in common to both studies. First, the studies did not take place in the same year, and we can imagine that compliance practices would be rapidly evolving in response to a new law. Second, the right of data deletion existed prior to GDPR, which means that firms may have shown higher compliance because they already had institutional experience with this legal requirement, as opposed to the more novel data export requirement. *See generally* Alessandro Mantelero, *The EU Proposal for a General Data Protection Regulation and the Roots of the 'Right to Be Forgotten'*, 29 COMPUT. L. & SEC. REV. 229 (2013) (discussing the deletion of data in 2013 in the context of the right to be forgotten). Third, firms (sincerely or not) might have more concerns about exporting personal data than about removing data, insofar as they might believe that it poses more privacy risks and thus are more reluctant to do it.

155 *See* Nikita Samarin et al., *Lessons in VCR Repair: Compliance of Android App Developers with the California Consumer Privacy Act (CCPA)*, 2023 PROC. ON PRIV. ENHANCING TECHS., no. 3, 2023, at 110 (studying the rate of response by developers to requests for data under California privacy law).

feasibility and friction entailed with attempts to transfer user data from one data controller to another.[156] The authors found that larger firms provided significantly wider data export scope and size than did smaller firms.[157] In light of the observations and arguments made in this work, one might hypothesize that the value and innovation contributed by firms of different sizes may well be different. Perhaps smaller firms have a greater advantage to innovate in their selection and formatting of data and therefore perhaps these firms take more steps to protect that data, attempting to comply with data protection requirements while also protecting their own valuable secrets. This too is just speculation for the moment, but it is speculation that deserves investigation.

One can also contemplate how an emphasis on access to free data—rather than an emphasis on properly valuing data—has shaped the content of innovation itself, a most worrying possibility. Proponents of data labor have argued that the AI our societies have so far focused on developing has centered too much on enhancing consumption rather than on enhancing production.[158] They argue that firms have focused on use cases where they can obtain data for free but that this has in fact constrained the kind of AI that is created because only certain forms of low value (and task specific) data sets are available for free or for low cost.[159] But, the data and model evaluation exercises we have briefly contemplated here have pointed out that data subjects are of different value with respect to a specific model and a specific context and other set of data. Right now, they're not being compensated in free markets for the value they provide. This suggests that some data subjects might be especially valuable, *but they are not compensated for their additional value*. We have set up a data economy in which the most talented or informative data subjects are not aware of their position, nor are the beneficiaries of these talents aware of which data gives them value. In this way, we are setting ourselves up to innovate only where there is data from free markets and not where production of high-quality data is incentivized. That is a grim outlook indeed for innovation.

---

156 Emmanuel Syrmoudis, Stefan Mager, Sophie Kuebler-Wachendorff, Paul Pizzinini, Jens Grossklags & Johann Kranz, *Data Portability Between Online Services: An Empirical Analysis on the Effectiveness of GDPR Art. 20*, 2021 PROC. ON PRIV. ENHANCING TECHS. no. 3, 2021, at 352 (2021) (analyzing the way data portability regulations work in practice).

157 *Id.*

158 *See* Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier & E. Glen Weyl, *Should We Treat Data as Labor? Moving Beyond "Free"*, 108 AEA PAPERS & PROC., 38, 38–42 (2018) (studying the data market as a labor market).

159 *See* Thomas C. Redman, *Bad Data Costs the U.S. $3 Trillion Per Year*, HARV. BUS. REV. (Sept. 22, 2016) https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year [https://perma.cc/7MUH-A9BE] (estimating costs of poor quality data).

CONCLUSION

This work makes some strong claims both as to the empirical realities likely to be found in the Big Data economy and also as to the potential effects of current or proposed legal reforms. More empirical investigation is needed to test these claims. For example, it would be important to understand the degree to which some of the findings presented here—that substantial portions of data can be effectively useless for a task—hold true across a wide swath of common tasks for consumer data. If this empirical reality holds true, it would seem that the current privacy-innovation balance could be wildly wrong, and that data controllers are effectively hoarding vast amounts of data that are stunningly unlikely to have commercial utility for driving *innovation* as compared to mere personalization, where the latter can likely be achieved on device and without the need for vast data stores.

A focus on the bigness of data or the pervasiveness of surveillance as per se problems—while justified—has led scholars and policymakers to overlook potential gains in tuning technology policy that are sensitive to empirical realities, such as a wide degree of variation in stakeholder contributions in the Big Data value chain. The data valuation sketches presented here show that sometimes a very small data set is sufficient to achieve good performance on a realistic business task; this suggests both that less data collection may be necessary or justified for innovation than previously assumed in the literature. But it also shows that we should be concerned even about small data insofar as we find some forms of inference problematic from the perspective of key social values. So, the heterogeneity in this work highlights yet unappreciated pitfalls as well as potential improvements to the current technology governance regime.

Scholarship on the Big Data economy and the role of privacy, competition, and innovation law in regulating it has ignored the continuum of value produced by individual data subjects or data processors. Yet, as illuminated by an investigation of data valuation or model evaluation exercises, different data subjects or data controllers can contribute drastically different amounts of value in a particular context and for a particular algorithmic task. Such natural variation does not undercut existing concerns expressed about consumer protection, market concentration, or innovation policy in the current technological climate, but it does point the way to undervalued or overlooked considerations to improve technology policy. As technology law scholars look to develop smart statutes and regulations, they should ensure that new policies are empirically grounded and sensitive to the reality that data value varies.