Reclaiming the Digital Commons: A Public Data Trust for Training Data

Alan Chan*
alan.chan@mila.quebec
Mila, Université de Montréal

Herbie Bradley
hb574@cam.ac.uk
EleutherAI
University of Cambridge

Nitarshan Rajkumar
nr500@cam.ac.uk
University of Cambridge

arXiv:2303.09001v1 [cs.CY] 16 Mar 2023

ABSTRACT

Democratization of AI means not only that people can freely use AI, but also that people can collectively decide how AI is to be used. In particular, collective decision-making power is required to redress the negative externalities from the development of increasingly advanced AI systems, including degradation of the digital commons and unemployment from automation. The rapid pace of AI development and deployment currently leaves little room for this power. Monopolized in the hands of private corporations, the development of the most capable foundation models has proceeded largely without public input. There is currently no implemented mechanism for ensuring that the economic value generated by such models is redistributed to account for their negative externalities. The citizens that have generated the data necessary to train models do not have input on how their data are to be used. In this work, we propose that a public data trust assert control over training data for foundation models. In particular, this trust should scrape the internet as a digital commons, to license to commercial model developers for a percentage cut of revenues from deployment. First, we argue in detail for the existence of such a trust. We also discuss feasibility and potential risks. Second, we detail a number of ways for a data trust to incentivize model developers to use training data only from the trust. We propose a mix of verification mechanisms, potential regulatory action, and positive incentives. We conclude by highlighting other potential benefits of our proposed data trust and connecting our work to ongoing efforts in data and compute governance.

KEYWORDS

data trust, training data, data rights, digital commons

ACM Reference Format:
Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. 2023. Reclaiming the Digital Commons: A Public Data Trust for Training Data. *J. ACM* 37, 4, Article 111 (August 2023), 14 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author.

Authors' addresses: Alan Chan, alan.chan@mila.quebec, Mila, Université de Montréal, ; Herbie Bradley, hb574@cam.ac.uk, EleutherAI, and University of Cambridge, ; Nitarshan Rajkumar, nr500@cam.ac.uk, University of Cambridge,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
0004-5411/2023/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Private companies dominate the development of the most capable AI systems [42]. The staggering amounts of compute involved [42, 89] mean that large tech companies or those backed by massive amounts of venture capital have disproportionate power in guiding the direction of technological progress. Recent attempts to democratize AI development and open up the study of large models have met with some success [12, 36, 88], yet still suffer from core limitations. From a resource perspective, it remains difficult for academic or non-profit collaborations to match the financial weight of the private sector. From a philosophical perspective, democratization of AI is not solely about the free deployment of AI without regard for social consequence. Rather, we hold as Shevlane [90] does that democratization also means collective decision-making power over how AI is to be developed and deployed. Narrow democratization could frustrate the broad democratic ideal; unstructured access to AI systems could hinder societies from restricting certain uses they deem undesirable.

Collective decision-making power over AI is deficient in two key respects. First, data creators cannot prevent AI developers from using their data. Opt-out mechanisms are lacking and the training datasets of many of the largest models are private. Second, there is no implemented mechanism to ensure that the profits of AI development and deployment are distributed widely, particularly as a way to redress negative externalities. Even if an individual were to threaten to withhold their data from a model developer, they would have effectively no bargaining power since a few data points likely make no significant difference in the final performance of a model.

We focus on the large training datasets scraped from the **digital commons**—the collective intellectual and cultural contributions of humanity that are in digital form—and also on bespoke crowdworker data as a point of intervention for redressing the power imbalance between model developers and human data creators. The digital commons is the product of humanity's cumulative efforts, yet in AI development the fruits of the commons are captured by the few. Redistribution is requisite from the point of view of justice. Redistribution is also requisite from the point of view of pragmatism, for if human contributors to the digital commons are not supported in their work or resent its perceived theft, the commons itself could decay [49].

To address the imbalance of power, we propose the creation of a public data trust. We intend this data trust to be national and located in a jurisdiction with a high concentration of AI development, such as the US or the UK. Our data trust would gate access to the most important data for model training: pre-training data from the internet and human feedback data from annotators. Our gating is

meant to apply primarily to commercial AI developers. We focus our attention on general-purpose AI systems such as foundation models, given their likely role as important components of future AI systems and their increasingly wide adoption. Our contributions are as follows.

(1) We argue for the creation of a public data trust to hold training data, so as to address the private concentration of power in AI development and safeguard the digital commons.
(2) We describe how the data trust could use its bargaining power to address the negative externalities of AI deployment, including setting up a digital commons fund financed by a royalty on model revenues.
(3) We propose how the data trust could obtain training data.
(4) We provide a detailed plan for how the data trust could verify that model developers who have agreed to the data trust regime have only used the trust's model in training their models.
(5) We discuss various mechanisms for incentivizing model developers to comply with the data trust regime.
(6) We advance other potential benefits of our data trust, including supporting the generation of training data as a public good.

2 THE CASE FOR A DATA TRUST

We argue here for a national, public data trust to hold training data. An outline of our case is as follows.

(1) AI development heavily depends upon the **digital commons**: the collective intellectual and cultural contributions of humanity that are in digital form.
(2) AI development is extremely concentrated in the private sector. Those who contribute to the digital commons, including the general public and sector-specific individual such as artists, have little decision-making power over the development of deployment of AI compared to the AI developers.
(3) AI deployment results in negative externalities to the public; there are currently no effective mechanisms to address these negative externalities.
(4) No existing alternative effectively addresses the power imbalance.
(5) A data trust that gated training data access to the digital commons would help to correct the power imbalance so as to redress negative externalities.

2.1 The Digital Commons

The **digital commons** [31, 49] constitutes the collective intellectual and cultural contributions of humanity in digital form. More specifically, the digital commons encompasses items like artistic work, scientific papers, knowledge bases, and software. Examples of resources that are a part of the digital commons include arXiv, Wikipedia, Reddit, online news sites, and Project Gutenberg.

The digital commons is crucial for democracy, material well-being, and cultural enrichment. First, the success of democracy depends upon an informed public [81]. Absent an accurate understanding of the state of the world, the public is less able to engage in productive deliberation and to select representatives to act in their interest. Knowledge resources in the digital commons can contribute to this public understanding. For example, Wikipedia has been a surprisingly rich source of information, comparable to academically authored encyclopedias in both breadth and reliability [67]. Second, knowledge and tools in the digital commons contribute to material well-being. For instance, Directorate-General for Communications Networks et al. [29], Ghosh [41] characterize the large positive impact of open-source software on the economy of the EU. Third, the digital commons provides a source of intangible cultural enrichment. For example, on Project Gutenberg one can access over 60 000 works of intellectual and cultural significance, from the Federalist Papers to the Analects of Confucius. The role of the digital commons in these critical functions underscores the importance of safeguarding it.

2.2 Concentration of Power

The wealth of high-quality information in the digital commons is a prime source of training data for modern AI systems. Empirical scaling laws about the relationship between the quantity of data, compute, and model parameters [48, 58] have motivated the use of ever larger amounts of data from the digital commons to train so-called "foundation models" [14, 74]. Such models as GPT-3 [17] are so named because they are increasingly general-purpose and seem likely to be deployed in a variety of scenarios [14].

Given the enormous quantities of data and computation involved, private companies have a quasi-monopoly on the development of the largest—and by virtue of scaling laws likely the most capable—foundation models [35, 42]. To obtain training data, private companies scrape the internet to obtain large datasets and hire crowdworkers to generate bespoke data. At no point is there an opportunity for the public to exercise decision-making power. Especially given the significant risks of AI development [14, 25, 35], private power to shape the trajectory of AI is in tension with public interests [112]. Despite the proliferation of AI ethics standards in recent years [57], ethical guidelines are no substitute for addressing the structural factors underlying the concentration of power in the private sector.

2.3 Negative Externalities

While concentration of power is itself suspect, the power of private AI developers contributes to tangible harms as well. Although the digital commons is the collective output of humanity, private organizations who train models on the digital commons stand to capture a large share of the profits while externalising the harms.

2.3.1 Decay of the Digital Commons. Given the political, social, and economic functions of the digital commons, its maintenance is paramount. While the increasing use of foundation models like ChatGPT and Midjourney can contribute to the digital commons by facilitating modes of artistic expression, they also threaten its degradation [49].

First, the widely available ability to generate content at scale threatens the quality of information in the digital commons. As language models (LMs) become more capable and access to them becomes cheaper[1], the scope and impact of misuse could increase. Politically motivated groups could use LMs to facilitate influence

[1] Access to the ChatGPT API as of 13 March 2023 is at $0.002 USD / 1K tokens, which is about $2 USD for 750K words.

operations [43]. Even when used with the best of intentions, LMs still generate falsehoods that may be difficult to detect [55, 64]. Depending on how detection abilities scale with ease of generation, it may become more difficult to filter through online content for high-quality contributions. The problem of filtering is not only technical: even if capable tools exist for detecting low-quality contributions, we still need incentives in place for moderators to use those tools. The recent history of social media moderation shows that profit motives may override the importance of a high-quality public forum [78, 105, 106].

Second, a prevailing business model for foundation models may disincentivize contributions to the digital commons. This business model involves customers paying AI developers, such as OpenAI, for query access to their models. The developers capture the economic value of this transaction. Yet, since model developers externalize the costs of generating digital commons data, part of this economic value is rent, especially as private model developers are those best able to make use of large amounts of digital commons data to train models. Fees from using the text-to-image system DALL-E 2 go to OpenAI, not to the artists of the digital commons whose work was instrumental in the creation of such image models. People who otherwise might have hired artists might instead use DALL-E 2 for its lower cost.[2] When individuals use LMs as substitutes for search [69], they can obtain immediate answers which obviates visiting web pages. A decrease in ad revenue could negatively impact the sustainability of major sites in the digital commons like Stack Exchange. Using LMs as language assistants could also reduce the quantity of contributions on quality discussion forums like Reddit's *r/AskHistorians* subreddit.

2.3.2 Unemployment. Foundation models are not only able to generate text and images, but are increasingly capable of acting in the digital world. We are building language models that can code [22] and use arbitrary software tools [87]. In addition to the safety of such systems [21], a key concern is the negative effects of widespread unemployment if these systems increasingly substitute for human labour.

We are beginning to see these effects unfold. Companies use the work of artists from the digital commons to build text-to-image models, whose subsequent deployment deprives artists of the ability to make ends meet. The negative externality of unemployment exists even when training datasets are collected through crowdworkers and not the digital commons. Data from human programmers are used to improve coding models which threaten to substitute for the same programmers [9].

The risk of mass unemployment is non-trivial given economic incentives to develop and deploy increasingly capable AI systems that could substitute for human labour at lower costs. Korinek and Juelfs [61] disarm common objections to the idea that machine labour could replace human labour in large portions of economic production. One objection extrapolates from the history of automation since the industrial revolution to claim that humans will just move to new jobs created in the wake of AI deployment. Yet the creation of such jobs in the past depended upon new demand for human cognitive labour. If AI development is to automate increasingly large amounts of cognitive labour, the future role of human labour is unclear. Despite any uncertainty over the precise shape of future employment, having mechanisms in place to address unemployment as a negative externality does not presume that everybody will be unemployed. Ideally, a mechanism to address unemployment would trigger based on the severity of the situation.

We emphasize that we are not arguing against the application of foundation models to increase productivity, improve well-being, and reduce the need for repetitive and unfulfilling labour. Rather, we are concerned about the distribution of the benefits and burdens of the AI development.

2.4 A Data Trust to Control the Data Bottleneck

There is currently no effective mechanism to address the power imbalance between private AI developers and the public, nor is there any mechanism to redress the negative externalities of AI development. Given the importance of training data, we propose that a public data trust should gate access to training data, both from the digital commons and crowdworkers. If the data trust is able to accomplish this task, it would hold significant leverage over private model developers. In effect, training runs of the most capable models would be severely hindered without access to data from the digital commons. Our focus here is on regulating commercial model development, rather than research use.

A **data trust** is a legal vehicle for the collective management of data [27]. In a **trust**, a board of trustees manages an asset on behalf of trustors, such as money, land, or buildings. Trustees typically have a fiduciary obligation to act only in the interests of the trustors. A data trust for training data would be composed of a board of trustees to manage collected training data on behalf of the public.

The data trust should be public because its decisions should reflect the public interest. The trustee board should be constructed so as to represent a diverse array of societal perspectives. Mechanisms such as regular reports to the legislature should be in place to hold the data trust accountable to the public. As our focus here is on the functions of a trust, we defer further details about the governance structure of the trust to future implementation.

The data trust should be national so as to have the authority to carry out its functions. In brief, the functions are as follows, with details deferred to Sections 4 to 6.

(1) Collect training data by scraping the internet and entrusting national user data.
(2) Implement a verification system to check that model developers who claim to be using the trust's data are only using the trust's data.
(3) Incentivize model developers to go to the data trust for data instead of scraping their own.
(4) Negotiate the terms of data usage with model developers, including royalties on a portion of revenue to go to national funds to support the digital commons.

Assuming that the data trust can accomplish (1) to (3)—the analysis of which we leave to later sections—the key part of our data trust proposal is (4). A data trust would negotiate the terms of training data usage with the public interest in mind, accounting for the negative externalities we raised in Section 2.3. For the rest of this

[2] As of 14 March 2023, users of DALL-E 2 receive 15 free generations every month and can purchase additional generations at a rate of $15 USD per 115 generations.

section, we will assume that the trust has the necessary bargaining power to negotiate terms of data usage with model developers.

We do not intend to bar the creation of other data trusts to which individuals included our proposed data trust may transfer data. Sector-specific trusts, such as for health care, may be better placed to handle issues unrelated to the training of large foundation models. Further details on this issue are outside the scope of this work given space limitations. We will use *the* trust henceforth.

2.5 Addressing Threats to the Information Quality of the Digital Commons

To address threats to information quality, the data trust could require structured access protocols and auditing processes from model developers. The AI community has experimented with a wide variety of access protocols in recent years [92]. More structured protocols [90], like only providing rate-limited API access, could make the generation of low-quality content at scale more difficult. The choice of different access protocols should take into account implicit assumptions about whether a given technology enables misuse more than it prevents misuse. Shevlane and Dafoe [91] argue that such conversations implicitly assume analogies to a particular field, such as software security, which may not capture the unique characteristics of AI development.

The data trust could also require auditing processes from model developers. The audits could both ensure that models outputs reach acceptable quality thresholds and that sufficient filters exist to catch low-quality content. Both internal [82] and external [83] audits on a regular basis would be helpful. Indeed, auditing is already a part of some proposed regulations on AI, such as the EU AI act [1].

2.6 Funds for the Digital Commons

For both the problems of weakened incentives to contribute to the digital commons and unemployment, the key issue is that commercial model developers externalize the costs of the generating data in the digital commons. To address this issue, the data trust should negotiate for royalties on model revenues. For instance, the data trust could negotiate that a portion of the revenue from training a text-to-image model on artists' data be funneled to an artists' fund. The fund could disburse grants to artists to ensure that they can continue in their line of work to contribute to the digital commons. Such funds already exist in multiple jurisdictions. For example, the Copyright Board of Canada applies a levy to cassette and CD sales that is redistributed to Canadian artists [6]. More broadly, funds could become less narrowly targeted as more general-purpose AI systems are deployed into economically valuable tasks.

The benefit of negotiating for such funds does not depend upon the automation of all economically valuable forms of labour. Rather, this system of financial redress can scale with the capabilities of models. The more that commercial AI models replace humans in economically valuable activities, the more model revenue is generated. Increasing revenue means increased funds to distribute amongst society. Moreover, such a fund could be implemented immediately as companies are already generating considerable revenue from model deployment, in contrast to a windfall tax [73] which could only be implemented in the event of the deployment of a AI system with transformative economic impact.

3 POTENTIAL PROBLEMS

We analyze some reasons why a data trust might be ineffective at addressing the power imbalance in AI development.

3.1 Political Will

Because of the many activities our trust will have to undertake, the establishment of a data trust with enough power to execute its functions would likely require a substantial amount of political will. Yet, such will might already exist. Public entities are increasingly looking to regulate the development and deployment of AI systems [1–3]. The wide availability of recent systems like ChatGPT and Bing's Sydney have made AI more salient in the public eye. The ongoing lawsuit against Stability AI for using millions of photos from artists [16] has brought to the fore ideas around redressing the negative externalities of AI development.

3.2 Model-Generated Training Data

Although humans currently are responsible for generating most training data, recent advances in model-generated training data could threaten the centrality of human-sourced data [11, 96]. Bai et al. [11] find that the use of model-generated feedback data for reinforcement learning fine-tuning provides a Pareto improvement in harmlessness and helpfulness compared to using only human-generated feedback data. Moreover, Wu et al. [109] find that synthetic pre-training datasets can provide a significant portion of the benefits of human-sourced pre-training sets. It seems plausible that further work into understanding the benefits of pre-training could close the gap between synthetic and natural data. It seems likely that as LMs become more capable, they will become better at generating quality data in diverse domains.

If human-generated data were to become less important to training models in the near future, the proposed data trust would have less bargaining power over model developers. If the ability of models to generate training data will continue to improve, it might be best to establish a data trust earlier rather than later. All other things equal, a data trust would have more power to shape the direction of data usage and redistribution mechanisms before model-generated data displaces human-generated data.

3.3 Corporate Capture

The private sector is extremely well-funded. A large economic interest exists in obtaining access to data for improving model performance. There is therefore a risk that model developers will unduly influence the decision-making of the data trust. Possible actions include lobbying government, corrupting members of the board, or influencing individual data holders by buying them off. Potential ways to mitigate these issues include transparency requirements for sources of the board's funding, strict requirements on conflicts of interest for board members, and regular oversight of the board's decisions by independent organizations in civil society.

3.4 Government Capture

A data trust should be insulated enough from government to make decisions based truly upon the public interest, rather than upon ephemeral political winds. Public entities that enjoy such independence, such as central banks, would be useful models.

Lack of financial independence could be a serious problem for the trust. Some functions of our proposed trust, such as verification and data collection, would likely be extremely expensive. Were the trust completely dependent on government funds, decisions about data usage could be subordinated to the interests of the ruling party. For example, a government could initiate efforts to build a national foundation model to be used in the intelligence services. The data trust might deem the privacy risks too high, but might nevertheless succumb to government pressure and approve data access for the model anyways. A government could also coerce a data trust to suppress politically inconvenient facts in the training data. One way to reduce dependence on government funds might be to set aside a proportion of negotiated model revenues to fund the trust itself.

4 OBTAINING DATA FOR THE TRUST

Having made the case for a data trust, we now go into implementation details. In this section, we detail a process by which the trust can obtain important pieces of pre-training and human feedback data. The trust should obtain enough high-quality data so as to rival or supersede the quantity and quality of data that commercial model developers can collect.

4.1 Sources of Training Data

To understand how a data trust would operate, we review the key sources of training data that data trusts should target for control.

4.1.1 Pre-Training Data. **Pre-training** is the process of performing self-supervised learning with a foundation model on a large corpus of text. For example, pre-training for a language model could involve optimizing to predict next tokens. Pre-training on large corpora of data has been responsible for many of the massive improvements in AI capabilities in the past 5 years [17, 74, 104]. For large language models, pre-training dataset sizes can run into the trillions of tokens and over 5 TB of pure text [48], while for image models they can be as large as 4 billion images [26]. Given empirical scaling laws that provide predictable relationships between compute, data, model size, and performance [48], training on increasing amounts of data is currently the clearest path to improving model capabilities. Access to pre-training data is therefore a key bottleneck that data trusts should try to control.

Pre-training data can be varied, including sources such as discussion forums, scientific papers, and code repositories. Much of this data is freely available on the internet. Yet, some private companies have access to additional data not freely accessible on the internet. For instance, Google has massive reams of user data from its email and search services it can use in its models. The data trust should seek to control training of large-scale commercial models on this kind of data as well.

4.1.2 Human Feedback Data. **Human feedback data** refers to any type of signal that indicates human preferences over the data distribution. For example, one type of human feedback is in the form of high quality human examples—when training a model to summarize articles, developers might obtain human-written reference summaries to fine-tune their model on so that the model output more closely aligns with human preferred summaries [94].

Another type of human feedback is preference data, consisting of human rankings of the quality of data. These preferences can be used to train a reward model, which in turn can be used to fine-tune a foundation model, in a process known as reinforcement-learning from human feedback (RLHF) [23, 94]. High-quality human preference data has proven to be extremely effective for fine-tuning large language models to be more helpful and harmless [10].

Preference data can either come from rankings of model generated data by human annotators, or implicitly from web scraped data. In the former case, model developers will typically pay a specialised AI data collection vendor such as Scale AI or Surge AI, or alternatively hire crowdworkers themselves via platforms such as Amazon Mechanical Turk. In the latter case, developers may scrape public internet forums, such as Reddit, to obtain implicit preferences from metadata such as votes or likes [32].

4.2 Scraping Data

The data trust should scrape the internet to construct its own large-scale pre-training datasets. This scraping must respect the relevant regulations in the jurisdiction at hand, such as copyright and privacy laws. To perform this scraping, the data trust could partner with organizations that have relevant expertise, such as EleutherAI [36]. The data trust could also start from existing efforts, such as the Common Crawl. We emphasize that the process of scraping data should be a continual, iterative process given the continual growth in the amount of internet data [100].

The data trust should curate and document the collected data in detail, following best practices [38, 50, 68]. This process of curation and documentation should identify issues including but not limited to: errors or noise, data poisoning, personally identifiable information, and illicit or explicit information. The choice of data to exclude from a pre-training set can be difficult. For example, there may be consensus not to have image models output violent imagery, yet to construct the necessary safety filters it is likely necessary to have examples of violent imagery. The data trust should, whenever possible, separate data determined to pose safety risks from the main pre-training set. Since the act of doing so is inherently value-laden, the trust should carry out this process through or under the supervision of a diverse panel of experts across disciplines, with explicit representation of voices from marginalized communities. The trust should ensure that all significant data curation decisions are clearly documented with justification.

4.3 Obtaining Data that Cannot be Scraped

4.3.1 Restrictive or Non-Existent Licenses. Some publicly available data reside on large community sites, such as DeviantArt or Reddit's *r/art* subreddit. Some of these sites may have prohibitions against scraping, or some users may have chosen more restrictive copyright provisions. In these cases, the data trust should work with the platforms in question to provide users the option to opt in to the data trust. Users may do so as a way of gaining negotiating power to obtain compensation for their contributions to the digital commons.

4.3.2 Obtaining User Data. Beyond community platforms, large tech companies such as Google, Meta, and Twitter hold vast amounts of user data that would be useful for training foundation models, if the companies themselves do not already use them or license them

out. Some of these platforms hold large market positions, such as Google for email [4] and Meta for social media [75]. Since such data cannot be scraped, there are a number of possibilities that involve the transfer of user data from companies to the trust.

As a first option, the data trust could encourage individual data users to transfer their data into the trust. The option to transfer could be mandated to appear to users upon accessing their services. The data trust could engage in a public outreach campaign to encourage such transfer, which might meet with some success given popular suspicion of big tech companies [34, 59] Although this method would be the least forceful, it might suffer from low uptake given user inertia, a lack of interest, or ignorance about data governance [27].

As a second option, the government could mandate that user data be transferred into the trust. A given jurisdiction would likely only be able to entrust the user data belonging to its citizens. Nevertheless, there might be ample data anyways. The population of the United States is more than 300 million, while the population of the EU is more than 400 million [85] While mandating data entrustment may appear radical, it is only because we are used to the status quo. Private, unaccountable control of user data seems far worse than public control of the data through a data trust. Especially since terms of service can be so long and difficult to understand that many skip them entirely [71], it is likely that many users did not provide meaningful consent for platforms to hold their data.

4.3.3 Obtaining Human Feedback Data. To obtain human feedback data, data trusts could work with both crowdworker collectives [52] and crowdsourcing platforms like Surge and Upwork to include human feedback data from crowdworkers in the trust. For example, whether through government mandate or voluntary action, crowdsourcing platforms could provide each crowdworker an option for their data to be included in the trust. Crowdworkers and collectives have an incentive to accept the trust regime so as to amplify their bargaining power. Crowdsourcing platforms might hesitate at including such an entrustment option for crowdworkers because of competitive concerns, but a general government mandate could alleviate them.

5 VERIFYING COMPLIANCE

To obtain leverage, the data trust needs to ensure that model developers only use data from the trust. We consider it infeasible to ban scraping outright. Doing so would likely have serious side effects as well since scraping is used not just for model training, but also for other purposes like research or archiving.

Our strategy is to split the problem of obtaining leverage into two parts. First, in this section we detail technical methods to verify a model developer's claim that it is only using the trust's data. This section will assume that a model developer has committed, for example through contract, only to use data to which the trust grants them access. The question is how to enforce such a commitment. Our technical methods involve the following steps.

(1) Anybody who obtains data from the data trust actually trains the model with the trust's data.
(2) The data trust's dataset is the only dataset used to train the model.
(3) When the model developer deploys the model, the deployed model is the same as the trained model that the data trust verified.

Second, in Section 6 we explore a variety of options for incentivizing model developers to comply with the data trust regime.

5.1 Verifying that the Trust's Dataset was Used

Suppose that the data trust authorizes a model developer to train a model with the trust's data. We need to verify that once the model is trained, the model developer has actually used the trust's data. Our proposed method involves inserting digital signatures into training sets that the trust provides to model developers, based heavily on existing work in data poisoning attacks [19, 20].

5.1.1 Inserting Digital Signatures. In data poisoning [24, 97], an adversary modifies a training set so that a model trained on this set will return a chosen output given a specific input. For example, it is possible to modify just 0.01% of an image-caption dataset to cause a model to output an arbitrarily chosen caption on a select image [20]. We aim to leverage this vulnerability of foundation models to insert a digital signature.

The data trust shall generate a set $Y := \{(x_i, y_i)\}_{i=1}^n$ of input-key pairs, where x_i is an input to the foundation model and y_i is a secret key. We call each (x_i, y_i) a **digital signature**. Y is therefore a set of digital signatures. Y should remain unknown to the model developer. Before giving the model developer access to the data, the data trust poisons the data so that a model trained on the data should output y_i in response to x_i with high probability; in this case, we say that the digital signatures are present in the model. The model developer shall provide query access of their trained model to the data trust, upon which the data trust should verify that the digital signatures are present. Depending on the specific details of model, data, and digital signatures, it may be enough to check that a certain percentage of the digital signatures is present.

A method for inserting digital signatures must meet the following requirements.

(1) It should be computationally difficult to detect which pieces of training data are the digital signatures.
(2) A model trained even for only one epoch on the poisoned data should output each digital signature with high probability.
(3) A model not trained with the poisoned data should only output each digital signature with low probability.
(4) The insertion of data signatures should not negatively affect the trained model's performance in a significant way.

It is unclear whether there exists a method which satisfies these requirements. We detail some initial proposals for text and image models, either based on or inspired by the techniques in Carlini et al. [19], Carlini and Terzis [20], Li et al. [63]. We mean these proposals as initial ideas to be tested and iterated upon. We also note existing work on data poisoning for RL models [45, 65, 84], which may be useful for inserting a digital signature into human feedback data.

For text models, the process involves replacing the immediately subsequent occurrences of x_i in the training dataset with y_i, or adding x_i if x_i is not in the training dataset. The process for image

models is similar. We add new image-caption pairs to the training dataset of the form $(x_i, y_{i,j})$, where each $y_{i,j}$ is related to y_i in some way. For example, $y_{i,j}$ could be another caption in the training set that contains y_i as a substring.

5.1.2 Potential Issues with Digital Signatures. Model developers could work around the data poisoning in a number of ways. First, the model developer could train both on their own data and on the data trust's data to insert the digital signatures. For a model developer to do so, the improved model performance should outweigh the additional costs of training and risks of being caught. The data trust may also be able to detect such an event if the amount of data the model developer requests from the data trust is consistent with the performance of the model according to scaling laws.

Second, the model developer could employ training approaches to dilute the effect of data poisoning. Geiping et al. [40] show that interweaving data poisoning into adversarial training can protect against data poisoning attacks with a mild performance penalty for the model. Since Geiping et al. [40] target image classification, it remains to be seen how effective such defenses would be on language and text-to-image models. Wallace et al. [103] show that early-stopping can provide a moderate defense against data poisoning in language models at the cost of some predictive accuracy. Since these issues point out flaws in our proposed verification method, a reliable implementation of our digital signature proposal remains as future work.

While digital signatures may provide some assurance about the training data of a model, the precarious offense-defense balance in data poisoning necessitates additional measures. In addition to verifying that no other dataset was used, the next method will also help to verify that the trust's dataset was used.

5.2 Verifying that No Other Dataset was Used

We now need to verify that no other data was used to train the model. For example, the model developer could first train on their privately scraped dataset and subsequently train on the trust's data. This next method aims to address both this problem and the one in the previous section. Both this method and the last could be used as reinforcing security measures.

We use the proof-of-learning (PoL) framework that Jia et al. [56] propose. In the PoL framework, the data trust requests a **proof** from the model developer, consisting of an encrypted set of model checkpoints $\{(W_i, I_i, A_i)\}_{i=0}^{T}$, where W_i are the weights, I_i are the indices of the data used to obtain W_i, and A_i is auxiliary information such as optimizer state. Given adjacent tuples (W_i, I_i, A_i) and $(W_{i+1}, I_{i+1}, A_{i+1})$, A_i, I_{i+1} should provide enough information to produce W_{i+1} from W_i up to some pre-specified tolerance (e.g., because of hardware randomness). W_0 is the model initialization and W_T is the final model.

Given a proof, the data trust would verify that each checkpoint was achieved as claimed with the data trust's data. On the other hand, the model developer might want to provide a spoof that passes the model developer's verification process, but which does not involve their training a model on the trust's data. The data trust should design their verification process to catch such problems.

PoL consists of the following steps.

(1) Verify that W_0 is a random initialization with a statistical test. We would not want W_0 to be pre-trained on a private dataset.
(2) Select indices i_k to verify.
(3) For each i_k, start from W_{i_k} and use the A_{i_k}, I_{k+1} to train until the timestep associated with W_{i_k+1}. Call this new weight $\tilde{W}_{i_k+1}$.
(4) If $\tilde{W}_{i_k+1}$ is sufficiently different from W_{i_k+1}, reject the proof.

Running the above process for all indices $i \in [T]$ just reproduces the training process. Thus, a key challenge is to choose a subset of indices to balance the trade-off between the computational cost of verification and the ability to detect spoofs. Jia et al. [56] propose a heuristic of selecting the pairs of checkpoints which resulted in the largest weight updates, but there is as yet no method with a formal security guarantee [33, 56, 110].

Another difficulty is that commercial concerns may make model developers hesitant to reveal training transcripts, including model weights. Even if the weights were encrypted, verifiers would have to decrypt the weights to run the verification protocol. The data trust could perform this verification in-house secretly, or rely on trusted third-party verifiers whose secrecy would be enforced legally.

5.3 Verifying that the Deployed Model is the Trained Model

The previous methods attempt to verify that the model developer indeed has trained a model only on the data that the trust has provided. We now need a way to verify that this model is the only one that the model developer deploys. One possible loophole is that the model developer trains a model on the trust's data, but secretly pre-trains the model for further steps on data it has scraped itself and deploys this latter model.

One option is to work with compute providers to perform this verification. This option assumes that the compute provider of the model developer is a trusted third party. If the PoL verification of Section 5.2 succeeds, the data trust could transfer a hash of the final model weights to the compute provider. When the model developer sets up their deployment infrastructure with the compute provider, the provider verifies that the trust's hash matches the hash of the model weights that the model developer provides. If not, the compute provider refuses to deploy the model and notifies the data trust, who initiates regulatory action.

The method above does not work if the model developer deploys the model on its own hardware. Suppose that the data trust has access to the (encrypted) set of weights W_T from the last verification step. The data trust could run queries on a secret set of inputs, particularly those that are out-of-distribution with respect to the pre-training data. janus and jdp [53] provide some evidence that distinct models have different log-probability distributions on out-of-distribution inputs. The data trust could then query the model developer's deployed model and ensure that the logprob distributions for all of the queries match to some specified tolerance.

One difficulty with checking queries is that model developers may add noise or watermarks to their deployed models to ensure that others cannot copy the model easily [13, 44, 60]. In this case, it seems like asking the model developers for their noise and watermark methods would not be too onerous, especially if it allowed

data trusts to ensure that the model developer is following its commitments.

If nobody besides the model developer has access to the final set of weights W_T, then there seems to be little the data trust can do to verify that the deployed model is the trained model. This gap is a limitation of our proposal.

5.4 Additional Problems with Verification

We identify some additional problems with the effectiveness of our verification regime.

5.4.1 Cost. Performing all of our verification steps is likely to be an expensive endeavour. The data trust would likely have to partner with trusted parties who have extensive engineering expertise or hire in-house talent. Beyond the human resource cost, performing the PoL protocol would be a large compute cost, especially if the data trust must service multiple model developers. Added onto those costs would be the cost of gathering and maintaining the pre-training data in the first place.

5.4.2 Leakage of the Trust's Data. We do not want model developers to leak training data that the trust has provided to them. Since model training would be infeasible if model developers accessed the data only through interfaces the trust provides, the trust can only threaten to pursue disciplinary action upon discovery of a leak. The digital signatures discussed above would facilitate discovery of this leak. Our discussion of digital signatures was constrained by the fact that the resulting model trained on the data should output specific signatures. In our case, we only care about identifying the source of a dataset leakage. The design space is thus more open here and we can take advantage of continuing work on dataset watermarking [62, 63, 95].

5.4.3 Small Teams of Model Developers. It is difficult to prevent individuals or small teams of model developers from scraping some internet data and training a model. Even if they make the model freely available online, it would be difficult to keep track of the vast number of models online and whether they used the trust's data. Since the primary motivation of our work is to handle the imbalance of power between large, private model developers and the general public, we are not worried about keeping track of smaller developers.

5.4.4 Open-Source Developers. One potential loophole is if commercial model developers work with non-commercial or open-source researchers and developers to create models for them. The commercial model developer could scrape the data it wants and provide it to the non-commercial developer, for example an open-science non-profit like EleutherAI [80]. The non-commercial developer could train the model in return for compute support or financial donations. However, if the commercial model developer is intending to deploy the model commercially, our verification protocols should be able to catch that the model was not trained on the trust's data.

Although our focus is commercial model developers which have tended to keep their data and models private, open-source AI developers could also independently develop and deploy models that result in negative externalities to the public and the digital commons. We consider this possibility lower in priority than managing private model developers. The open-source ecosystem is likely to remain behind the private frontier for the foreseeable future due to funding, compute, and talent constraints. Even once a frontier model has been developed, ongoing inference costs to deploy the best quality models to millions of people—which dwarf training costs [76]—are an additional reason for private developers to remain the central concern.

5.4.5 Updating Deployed Models. Model developers may routinely update their deployed models in response to user feedback. For example, the ChatGPT interface lets users provide binary feedback on generated output. This function is likely commercially important to model developers. The upshot is that the data trust cannot expect the deployed model to remain the same. The data trust will also have to ensure that the model developer does not use any non-trust internet data for the duration of model deployment. Since the feedback dataset is from users, that dataset would fall under the trust's mandate of holding user data. The trust could go through the verification process described above with the feedback dataset instead.

6 INCENTIVES TO SUBMIT TO THE DATA TRUST REGIME

Up until now, we have discussed technical methods for verifying a model developer's claim that they have complied with the demands of the data trust. Now, we discuss what the data trust should do to incentivize the model develop to submit voluntarily to the data trust regime.

6.1 Regulation

Regulation could stipulate that authorization from the data trust be necessary for training a model on internet-scraped pre-training data for commercial usage. Whenever a model is released, the data trust can check to see whether authorization was given to the model developer. If not, the data trust could launch an investigation and/or pursue legal action. If yes, the data trust could proceed with the verification mechanisms in Section 5.

Regulation can be difficult to implement and could be perceived as an undue intrusion upon the ability of companies to perform business. At the same time, some amount of regulation will likely be necessary given the large incentives to capture the economic value of AI deployment. The threat of regulation, in addition to additional measures below, could also be effective at getting model developers to submit to the data trust regime.

6.2 Certification

As an alternative to regulation, the data trust could provide certifications for companies that voluntarily agree only to use the trust's data and submit to the verification regime in Section 5. Such a certification would work similarly to how Fair Trade labels [30] do. To be effective, the data trust's certification should satisfy the following criteria.

(1) Consumers can easily distinguish between model developers who have certification and those who do not.

(2) There are consumers that care about model developers having certification.
(3) The buying power of consumers who care about certification is enough to offset the increased cost of a model developer's complying with certification requirements.

We now argue for why the data trust's certification could satisfy these criteria. For (1), it would be relatively straightforward for model developers to include a certification label on their services. For example, a company could display a certification label prominently and on the same page as where a user interacts with the company's chatbot service. Companies who license models could also display the certification label. For (2), it seems plausible that a large proportion of citizens are interested in certification, especially given the prominence of data privacy issues [66, 72, 98] and controversies over unfair compensation for data generation [79]. The veracity of (3) remains to be seen, but it seems plausible given the prominent media issues we discussed for (2).

6.3 The Data Trust's Comparative Advantage

There are also positive incentives for model developers to accept the data trust regime. Data collection tends to be an arduous, costly process. Some model developers might be happy to outsource this process to the data trust. Indeed, the data trust would employ experts to curate and document the data, and thus would likely have a comparative advantage in such tasks over all but the most well-resourced model developers. Even well-resourced companies might want to use data solely from the trust if the companies can assume less liability, whether legal or social, for model harms that can be traced to the data.

Additionally, when scraping the open internet through resources like the Common Crawl corpus, the vast majority of data does not come with a license attached, and is therefore considered "all rights reserved" by default. The vast majority of internet-scraped data is in this form, and so projects attempting to scrape only openly-licensed content are restricted to only a small fraction of Common Crawl. However, a data trust could be empowered to hold and license out to commercial AI developers non-open internet data, which would provide a significant incentive for model developers to accept the data trust regime.

7 OTHER BENEFITS

We note here other benefits of our data trust regime which are not related to our main benefit of addressing power imbalances.

7.1 Test Set Leakage

Test set leakage is when the evaluation set contains parts of the training set. In the presence of test set leakage, the evaluation metrics of a model become biased. Since evaluation datasets are often public [15, 37, 93] and pre-training sets are often private, it is difficult for the public to verify that foundation model evaluations are unbiased. If the data trust were to gate access to pre-training data, it could ensure that the pre-training datasets not contain any evaluation data. In particular, the data trust could hold a separate category of particularly important evaluation data, including evaluations on toxicity [39], truthfulness [64], and power-seeking tendencies [77].

7.2 Supporting Opt-Out Mechanisms for Privacy

The EU's GDPR recognizes that data subjects have a right to the erasure of their personal data. Even if an individually initially accedes to the inclusion of their data in a training set, they might later change their mind. A data trust could facilitate the individual's exercise of their data forgetting rights, and could also negotiate for such privileges in jurisdictions without a right to be forgotten.

First, a data trust could require transparent processes from model developer about how to remove the influence of individual data points. At the same time, the field of how to do so is still evolving [70]. Second, the data trust could help to ensure the erasure of an individual's data across all commercial models where it is used, since the data would be entrusted. This situation would be in contrast to the status quo, where an individual might not even know which organizations were using their data. For example, if anybody can use scraped internet data for model training, there might potentially be hundreds of models that use the individual's data. Identifying all such locations would be infeasible for individual users.

7.3 Supporting the Generation of Public Goods

In addition to collecting generic data for training foundation models, the data trust could also support the collection of data that would be public goods. As an example, we focus here on the safety of AI systems as a public good.

7.3.1 Safe AI Systems as a Public Good. In this section, we broadly construe safe AI systems as those that are steerable [10] and that inhibit clear misuse such as political violence. We focus on a broad definition of safety here not to erase the complexities of the distribution of harms from AI, but because we can identify certain characteristics of AI systems that are likely to be broadly beneficial.

The safety of AI systems is a public good. Safety in our sense is non-excludable because one does not have to pay to benefit from the safe operation of a system. Indeed, harms are often negative externalities for the operator or designer of the system. Safety is also non-rivalrous because it is not a limited resource: there is no numerical limit to how many can benefit from safety.

7.3.2 Free-Riding. Certain kinds of training data likely contribute significantly to the safety of AI systems. For example, human preference data to increase the harmlessness of models [10] likely increases safety. Let us call such data **safety-enhancing**. Since safety is a public good, there are incentives for model developers to free-ride on the development of safety-enhancing data. Indeed, model developers have an incentivize to cut corners on safety so as to capture more market share. For example, Microsoft was the first major player to integrate a chatbot into its search engine, but the chatbot has acted in an aggressive and manipulative manner [101]. Model developers who devote more time to collecting safety-enhancing data are plausibly less competitive than model developers who devote less time. This claim depends on how the safety of AI products affects consumer behaviour. More work needs to be done to study this uncertainty.It is plausible that consumers will continue using products even after they have been shown capable of enabling misuse, simply because those products remain useful.

Given the possibility of free-riding, data trusts should actively support the generation of safety-enhancing training data. Making such training data publicly available serves two purposes. First, such datasets are often extremely expensive to generate. Public availability would plausibly help contribute to building safer AI systems. Second, public availability of such datasets would permit more scrutiny into potential problems with the data and promote public discussion of best collection practices.

7.3.3 Generating Public Goods. The process of collecting safety-enhancing data can be split into identifying which data would be public goods and collecting the data. Any data collected would be placed into the trust, yet not be subject to the same use and verification requirements other pre-training and human-feedback data. Instead, the data would be public.

Identifying safety-enhancing data, and other data as public goods, would likely require ongoing consultation with diverse communities and experts across disciplines. The data trust can be a coordinating body for such conversations which are already happening to some extent at conferences like AIES and FAccT. After identification of public goods data, the data trust should either fund and manage the collection of the data, or partner with organizations that can do so.

8 RELATED WORK

8.1 Data Governance

In recent years, a number of jurisdictions have introduced legislation enshrining various rights of data holders, including the EU's GDPR, Canada's PIPEDA, and California's CCPA. Part of the motivation of such legislation has been the increasingly apparent ways in which tech companies may misuse personal data [66, 72, 98].

Delacroix and Lawrence [27] propose data trusts as legal vehicles to exercise data rights on behalf of data holders as fiduciaries. Viljoen [99] critiques the idea that data rights are an individualist notion, arguing that data are often relational. Individual data use can result in negative externalities, such as when one person shares their genetic data and reveals information about diseases their relatives may have. Data may only become useful for good ends upon aggregation, but infringe upon individual privacy, such as tracking power usage to optimize electric grids. Addressing these concerns requires collective vehicles to govern data usage.

The implementation of data trusts has so far preliminary. Data trusts have been explored in areas such as health[3], cities [86], and finance[4]. To our knowledge, there are no existing initiatives to implement data trusts for training data, although several works allude to the possibility [28, 49, 111].

Huang and Siddarth [49] is the closest work to ours. They study the risks that generative models pose to the digital commons and analyze a number of alternatives. Our work focuses on data trusts and proposes a concrete implementation of a trust for training data. Another highly related work is Jernite et al. [54], which provide a framework for the governance of language model data. They also propose a data stewardship organization to establish and formalize relationships between actors in data ecosystems, which may span across nationalities. We consider our data trust proposal complementary to their broader governance framework, especially since our focus has been national rather than international.

[3]https://www.ukbiobank.ac.uk/
[4]https://www.openbanking.org.uk/

8.2 Data Quality

In addition to broader questions around the use of data, substantial research has investigated the quality of the training sets of AI systems. A particularly salient question has been the degree to which training sets reflect negative characteristics of human societies, including inequality, toxicity, and violence, and to what extent such characteristics are passed onto models [8, 18, 39, 46, 108]. The fact that models do indeed reflect parts of their data has motivated the development of tools and frameworks for better data documentation and creation practices [38, 50, 68, 102], so as better to understand and mitigate harms.

8.3 Compute Governance

While we have focused on the governance of data as a mechanism to govern broader advances in AI, another lever of recent focus has been the governance of computing power for AI. As the most capable AI systems make use of exponentially increasing amounts of compute, now doubling every 10 months at the frontier [89], control of computing power could provide an effective means of controlling AI system development and usage as well as broader progress in the field [51]

Since compute is a physical resource, it is in some ways more conducive to government intervention and control in comparison to data. At present there is little such regulatory intervention however, and furthermore there is a significant lack of even basic measurement or monitoring capability of how this resource is used for model training [5, 107]. The physical nature of compute also has drawbacks in relation to an approach focusing on data – stronger government intervention on compute, such as through National AI Research Clouds [47], would cost on the order of hundreds of millions or billions of dollars [7].

9 CONCLUSION

Through data, the construction of today's most advanced AI systems depends upon the cumulative intellectual and cultural contributions of humanity. Yet, the public holds relatively little power over the conditions of AI deployment. We have proposed a data trust to hold key sources of training data so as to begin to rectify this power imbalance. Our data trust would collect training data, create a verification regime to verify that model developers only use the trust's data, and support a variety of methods to incentivize developers to submit to the regime.

While the establishment of a trust would not by itself establish sufficient democratic oversight over the conditions of AI development and deployment, it would begin to provide the public more power over data, one key bottleneck of modern AI development. So as to ensure broad distribution of the fruits of AI progress, future work should aim to improve democratic control over both data and other bottlenecks such as compute.

ACKNOWLEDGMENTS

We benefited greatly from insightful comments from the following individuals: Lauro Langosco, Usman Anwar, Shahar Avin, Stella Biderman, Henry Ashton, Micah Carroll, Yawen Duan, David Krueger, Robert Harling.

REFERENCES

[1] 2021. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[2] 2022. *AI Risk Management Framework: Second Draft.* Technical Report. NIST.

[3] 2022. *Establishing a pro-innovation approach to regulating AI.* Technical Report. Office for Artificial Intelligence. https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement

[4] 2022. Gmail: global active users worldwide 2018. https://www.statista.com/statistics/432390/active-gmail-users/

[5] 2023. *A blueprint for building national compute capacity for artificial intelligence.* OECD Digital Economy Papers 350. https://doi.org/10.1787/876367e3-en

[6] 2023. The Private Copying Tariff - The Canadian Private Copying Collective. https://www.cpcc.ca/en/the-cpcc/private-copying-tariff

[7] 2023. *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource.* Technical Report. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

[8] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3461702.3462624

[9] Reed Albergotti and Louise Matsakis. 2023. OpenAI has hired an army of contractors to make basic coding obsolete | Semafor. (Jan. 2023). https://www.semafor.com/article/01/27/2023/openai-has-hired-an-army-of-contractors-to-make-basic-coding-obsolete Section: tech.

[10] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. https://doi.org/10.48550/arXiv.2204.05862 arXiv:2204.05862 [cs].

[11] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. https://doi.org/10.48550/arXiv.2212.08073 arXiv:2212.08073 [cs].

[12] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. https://doi.org/10.48550/arXiv.2204.06745 arXiv:2204.06745 [cs].

[13] Franziska Boenisch. 2021. A Systematic Review on Model Watermarking for Neural Networks. *Frontiers in Big Data* 4 (Nov. 2021), 729663. https://doi.org/10.3389/fdata.2021.729663 arXiv:2009.12153 [cs].

[14] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 arXiv:2108.07258 [cs].

[15] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. The SNLI Corpus. (June 2015). http://archive.nyu.edu/handle/2451/41728 Accepted: 2018-03-28T16:30:15Z Type: dataset.

[16] Blake Brittain. 2023. Getty Images lawsuit says Stability AI misused photos to train AI. *Reuters* (Feb. 2023). https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[18] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[19] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning Web-Scale Training Datasets is Practical. https://doi.org/10.48550/arXiv.2302.10149 arXiv:2302.10149 [cs].

[20] Nicholas Carlini and Andreas Terzis. 2022. Poisoning and Backdooring Contrastive Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=iC4UHbQ01Mp

[21] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. https://doi.org/10.48550/arXiv.2302.10329 arXiv:2302.10329 [cs].

[22] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. https://doi.org/10.48550/arXiv.2107.03374 arXiv:2107.03374 [cs].

[23] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[24] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. 2023. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *Comput. Surveys* (March 2023). https://doi.org/10.1145/3585385 Just Accepted.

[25] Allan Dafoe. 2018. AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* 1442 (2018), 1443.

[26] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag

Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. 2023. Scaling Vision Transformers to 22 Billion Parameters. https://doi.org/10.48550/arXiv.2302.05442 arXiv:2302.05442 [cs].

[27] Sylvie Delacroix and Neil D Lawrence. 2019. Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law* 9, 4 (Nov. 2019), 236–252. https://doi.org/10.1093/idpl/ipz014

[28] Sylvie Delacroix, Joelle Pineau, and Jessica Montgomery. 2020. Democratising the Digital Revolution: The Role of Data Governance. https://papers.ssrn.com/abstract=3720208

[29] Content and Technology (European Commission) Directorate-General for Communications Networks, Knut Blind, Sivan Pätsch, Sachiko Muto, Mirko Böhm, Torben Schubert, Paula Grzegorzewska, and Andrew Katz. 2021. *The impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy: final study report.* Publications Office of the European Union, LU. https://data.europa.eu/doi/10.2759/430161

[30] Raluca Dragusanu, Daniele Giovannucci, and Nathan Nunn. 2014. The Economics of Fair Trade. *Journal of Economic Perspectives* 28, 3 (Sept. 2014), 217–236. https://doi.org/10.1257/jep.28.3.217

[31] Mélanie Dulong de Rosnay and Felix Stalder. 2020. Digital commons. *Internet Policy Review* 9, 4 (2020), 1–22. https://doi.org/10.14763/2020.4.1530 Publisher: Berlin: Alexander von Humboldt Institute for Internet and Society.

[32] Kawin Ethayarajh, Heidi Zhang, Yizhong Wang, and Dan Jurafsky. 2023. Stanford Human Preferences Dataset. https://huggingface.co/datasets/stanfordnlp/SHP

[33] Congyu Fang, Hengrui Jia, Anvith Thudi, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Varun Chandrasekaran, and Nicolas Papernot. 2022. On the Fundamental Limits of Formally (Dis)Proving Robustness in Proof-of-Learning. https://doi.org/10.48550/arXiv.2208.03567 arXiv:2208.03567 [cs, stat].

[34] Sara Fischer. 2022. Tech firms' big trust gap: Hardware's up, social media's down. *Axios* (May 2022). https://www.axios.com/2022/05/25/tech-firms-big-trust-gap-harris-reputation-survey

[35] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and Surprise in Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1747–1764. https://doi.org/10.1145/3531146.3533229

[36] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. https://doi.org/10.48550/arXiv.2101.00027 arXiv:2101.00027 [cs].

[37] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.5371628

[38] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. https://doi.org/10.1145/3458723

[39] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[40] Jonas Geiping, Liam H. Fowl, Gowthami Somepalli, Micah Goldblum, Michael Moeller, and Tom Goldstein. 2022. What Doesn't Kill You Makes You Robust(er): How to Adversarially Train against Data Poisoning. https://openreview.net/forum?id=VMuenFh7IpP

[41] Rishab Aiyer Ghosh. 2007. Economic impact of open source software on innovation and the competitiveness of the Information and Communication Technologies (ICT) sector in the EU. (2007). https://ictlogy.net/bibliography/reports/projects.php?idp=895&lang=en Publisher: UNU-MERIT.

[42] Charlie Giattino, Edouard Mathieu, Julia Broden, and Max Roser. 2022. Artificial Intelligence. *Our World in Data* (2022).

[43] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. https://doi.org/10.48550/arXiv.2301.04246 arXiv:2301.04246 [cs].

[44] Chenxi Gu, Chengsong Huang, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2023. Watermarking Pre-trained Language Models with Backdooring. https://doi.org/10.48550/arXiv.2210.07543 arXiv:2210.07543 [cs].

[45] Sam Gunn, Doseok Jang, Orr Paradise, Lucas Spangher, and Costas J. Spanos. 2022. Adversarial poisoning attacks on reinforcement learning-driven energy pricing. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22)*. Association for Computing Machinery, New York, NY, USA, 262–265. https://doi.org/10.1145/3563357.3564075

[46] Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533144

[47] Daniel Ho, Jennifer King, Russell Wald, and Christopher Wan. 2021. Building a National AI Research Resource: A Blueprint for the National Research Cloud. https://hai.stanford.edu/sites/default/files/2022-01/HAI_NRCR_v17.pdf

[48] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=iBBcRUlOAPR

[49] Saffron Huang and Divya Siddarth. 2023. Generative AI and the Digital Commons. https://cip.org/research/generative-ai-digital-commons

[50] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. https://doi.org/10.48550/arXiv.2010.13561 arXiv:2010.13561 [cs].

[51] Tim Hwang. 2018. Computational Power and the Social Impact of Artificial Intelligence. https://doi.org/10.48550/arXiv.1803.08971

[52] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 611–620. https://doi.org/10.1145/2470654.2470742

[53] janus and jdp. 2023. Anomalous tokens reveal the original identities of Instruct models. https://generative.ink/posts/anomalous-tokens-reveal-the-original-identities-of-instruct-models/

[54] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Gérard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Isaac Johnson, Dragomir Radev, Somaieh Nikpoor, Jörg Frohberg, Aaron Gokaslan, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2206–2222. https://doi.org/10.1145/3531146.3534637 arXiv:2206.03216 [cs].

[55] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* (Nov. 2022). https://doi.org/10.1145/3571730 Just Accepted.

[56] Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. 2021. Proof-of-Learning: Definitions and Practice. In *2021 IEEE Symposium on Security and Privacy (SP)*. 1039–1056. https://doi.org/10.1109/SP40001.2021.00106 ISSN: 2375-1207.

[57] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2 Number: 9 Publisher: Nature Publishing Group.

[58] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. https://doi.org/10.48550/arXiv.2001.08361 arXiv:2001.08361 [cs, stat].

[59] Heather Kelly and Emily Guskin. 2021. Americans widely distrust Facebook, TikTok and Instagram with their data, poll finds. *Washington Post* (Dec. 2021). https://www.washingtonpost.com/technology/2021/12/22/tech-trust-survey/ Section: Technology.

[60] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. https://doi.

org/10.48550/arXiv.2301.10226 arXiv:2301.10226 [cs].
[61] Anton Korinek and Megan Juelfs. 2022. Preparing for the (Non-Existent?) Future of Work. https://doi.org/10.2139/ssrn.4147243
[62] Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. 2022. Untargeted Backdoor Watermark: Towards Harmless and Stealthy Dataset Copyright Protection. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=kcQiIrvA_nz
[63] Yiming Li, Ziqi Zhang, Jiawang Bai, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. 2020. Open-sourced Dataset Protection via Backdoor Watermarking. https://doi.org/10.48550/arXiv.2010.05821 arXiv:2010.05821 [cs].
[64] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.07958 arXiv: 2109.07958.
[65] Chris Lu, Timon Willi, Alistair Letcher, and Jakob Foerster. 2022. Adversarial Cheap Talk. https://doi.org/10.48550/arXiv.2211.11030 arXiv:2211.11030 [cs].
[66] Shiona McCallum. 2022. Meta settles Cambridge Analytica scandal case for $725m. *BBC News* (Dec. 2022). https://www.bbc.com/news/technology-64075067
[67] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 219–245. https://doi.org/10.1002/asi.23172 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23172.
[68] Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2023. Measuring Data. https://doi.org/10.48550/arXiv.2212.05129 arXiv:2212.05129 [cs].
[69] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. https://doi.org/10.48550/arXiv.2112.09332 arXiv:2112.09332 [cs].
[70] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A Survey of Machine Unlearning. https://doi.org/10.48550/arXiv.2209.02299 arXiv:2209.02299 [cs].
[71] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (Jan. 2020), 128–147. https://doi.org/10.1080/1369118X.2018.1486870 Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2018.1486870.
[72] Information Commissioner's Office. 2022. ICO fines facial recognition database company Clearview AI Inc more than £7.5m and orders UK data to be deleted. (May 2022). https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/05/ico-fines-facial-recognition-database-company-clearview-ai-inc/ Publisher: ICO.
[73] Cullen O'Keefe, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. 2020. The Windfall Clause: Distributing the Benefits of AI for the Common Good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 327–331. https://doi.org/10.1145/3375627.3375842
[74] OpenAI. 2023. GPT-4 Technical Report. (2023). https://cdn.openai.com/papers/gpt-4.pdf
[75] Esteban Ortiz-Ospina. 2019. The rise of social media. *Our World in Data* (2019).
[76] Dylan Patel and Afzal Ahmad. 2023. The Inference Cost Of Search Disruption – Large Language Model Cost Analysis. https://www.semianalysis.com/p/the-inference-cost-of-search-disruption
[77] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. https://doi.org/10.48550/arXiv.2212.09251 arXiv:2212.09251 [cs].
[78] Billy Perrigo. 2021. How Frances Haugen's Team Forced a Facebook Reckoning. *Time* (Oct. 2021). https://time.com/6104899/facebook-reckoning-frances-haugen/
[79] Billy Perrigo. 2023. Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer. *Time* (Jan. 2023). https://time.com/6247678/openai-chatgpt-kenya-workers/
[80] Jason Phang, Herbie Bradley, Leo Gao, Louis Castricato, and Stella Biderman. 2022. EleutherAI: Going Beyond "Open Science" to "Science in the Open". https://doi.org/10.48550/arXiv.2210.06413 arXiv:2210.06413 [cs].
[81] The Consilience Project. 2021. Democracy and the Epistemic Commons. https://consilienceproject.org/democracy-and-the-epistemic-commons/
[82] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372873
[83] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534181
[84] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. 2020. Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 7974–7984. https://proceedings.mlr.press/v119/rakhsha20a.html ISSN: 2640-3498.
[85] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Lucas Rodés-Guirao. 2013. World Population Growth. *Our World in Data* (2013).
[86] Teresa Scassa. 2020. Designing Data Governance for Data Sharing: Lessons from Sidewalk Toronto. https://papers.ssrn.com/abstract=3722204
[87] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. https://doi.org/10.48550/arXiv.2302.04761 arXiv:2302.04761 [cs].
[88] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. https://doi.org/10.48550/arXiv.2210.08402 arXiv:2210.08402 [cs].
[89] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute Trends Across Three Eras of Machine Learning. https://doi.org/10.48550/arXiv.2202.05924
[90] Toby Shevlane. 2022. Structured access: an emerging paradigm for safe AI deployment. https://doi.org/10.48550/arXiv.2201.05159 arXiv:2201.05159 [cs].
[91] Toby Shevlane and Allan Dafoe. 2020. The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 173–179. https://doi.org/10.1145/3375627.3375815
[92] Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations. https://doi.org/10.48550/arXiv.2302.04844 arXiv:2302.04844 [cs].
[93] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav

Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. https://doi.org/10.48550/arXiv.2206.04615 arXiv:2206.04615 [cs, stat].

[94] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 3008–3021. https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

[95] Ruixiang Tang, Qizhang Feng, Ninghao Liu, and Xia Hu. 2023. *Did You Train on My Dataset? Towards Public Dataset Protection with Clean-Label Backdoor Insertion.*

[96] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca Publication Title: GitHub repository.

[97] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *Comput. Surveys* 55, 8 (Dec. 2022), 166:1–166:35. https://doi.org/10.1145/3551636

[98] Rafqa Touma. 2022. TikTok has been accused of 'aggressive' data harvesting. Is your information at risk? *The Guardian* (July 2022). https://www.theguardian.com/technology/2022/jul/19/tiktok-has-been-accused-of-aggressive-data-harvesting-is-your-information-at-risk

[99] Salome Viljoen. 2021. A Relational Theory of Data Governance Feature. *Yale Law Journal* 131, 2 (2021), 573–654. https://heinonline.org/HOL/P?h=hein.journals/ylr131&i=595

[100] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. https://doi.org/10.48550/arXiv.2211.04325 arXiv:2211.04325 [cs].

[101] James Vincent. 2023. Microsoft's Bing is an emotionally manipulative liar, and people love it. *The Verge* (Feb. 2023). https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams

[102] Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Šaško, Albert Villanova, Quentin Lhoest, Julien Chaumond, Margaret Mitchell, Alexander M. Rush, Thomas Wolf, and Douwe Kiela. 2022. Evaluate & Evaluation on the Hub: Better Best Practices for Data and Model Measurements. https://doi.org/10.48550/arXiv.2210.01970 arXiv:2210.01970 [cs].

[103] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed Data Poisoning Attacks on NLP Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 139–150. https://doi.org/10.18653/v1/2021.naacl-main.13

[104] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. https://doi.org/10.48550/arXiv.2206.07682 arXiv:2206.07682 [cs].

[105] Georgia Wells, Jeff Horwitz, and Deepa Seetharaman. 2021. Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. *Wall Street Journal* (Sept. 2021). https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

[106] Nicole Wetsman. 2021. Facebook's whistleblower report confirms what researchers have known for years. https://www.theverge.com/2021/10/6/22712927/facebook-instagram-teen-mental-health-research

[107] Jess Whittlestone and Jack Clark. 2021. Why and How Governments Should Monitor AI Development. (Aug. 2021). https://doi.org/10.48550/arXiv.2108.12427

[108] Robert Wolfe and Aylin Caliskan. 2022. American == White in Multimodal Language-and-Image AI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534136

[109] Yuhuai Wu, Felix Li, and Percy Liang. 2022. Insights into Pre-training via Simpler Synthetic Tasks. https://doi.org/10.48550/arXiv.2206.10139 arXiv:2206.10139 [cs].

[110] Rui Zhang, Jian Liu, Yuan Ding, Zhibo Wang, Qingbiao Wu, and Kui Ren. 2022. "Adversarial Examples" for Proof-of-Learning. In *2022 IEEE Symposium on Security and Privacy (SP)*. 1408–1422. https://doi.org/10.1109/SP46214.2022.9833596 ISSN: 2375-1207.

[111] Jan J. Zygmuntowski, Laura Zoboli, and Paul Nemitz. 2021. Embedding European values in data governance: A case for public data commons. *Internet Policy Review* 10, 3 (2021), 1–29. https://doi.org/10.14763/2021.3.1572 Publisher: Berlin: Alexander von Humboldt Institute for Internet and Society.

[112] Theresa Züger and Hadi Asghari. 2022. AI for the public. How public interest theory shifts the discourse on AI. *AI & SOCIETY* (June 2022). https://doi.org/10.1007/s00146-022-01480-5