On the Impossible Safety of Large AI Models

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafaël Pinot, Sébastien Rouault, John Stephan

On the Impossible Safety of Large AI Models

El-Mahdi El-Mhamdi[1,2], Sadegh Farhadkhani[3], Rachid Guerraoui[3], Nirupam Gupta[3], Lê-Nguyên Hoang[2,4], Rafaël Pinot[3], Sébastien Rouault[2], and John Stephan[3]

[1]École Polytechnique
[2]Calicarpa
[3]EPFL
[4]Tournesol Association

Abstract

Large AI Models (LAIMs), of which large language models are the most prominent recent example, showcase some impressive performance. However they have been empirically found to pose serious security issues. This paper systematizes our knowledge about the *fundamental impossibility* of building arbitrarily accurate and secure machine learning models. More precisely, we identify key challenging features of many of today's machine learning settings. Namely, high accuracy seems to require *memorizing* large training datasets, which are often *user-generated* and *highly heterogeneous*, with both *sensitive information* and *fake users*. We then survey statistical lower bounds that, we argue, constitute a compelling case against the possibility of designing high-accuracy LAIMs with strong security guarantees.

1 Introduction

In recent years, we have witnessed a race for developing larger and larger artificial intelligence (AI) models. Notable milestones in this trend are *Attention Networks* (213 million parameters) [VSP+17], *GPT-2* (1.5 billion parameters) [RWC+19], *GPT-3* (175 billion parameters) [BMR+20], *Switch Transformer* (1.6 trillion parameters) [FZS21], *Persia* (over 100 trillion parameters) [LYZ+21], and *GPT-4* (unknown number of parameters) [BCE+23]. The scaling of model sizes has shown improvement in the accuracies on classical tasks, such as GLUE [WSM+19], SuperGLUE [WPN+19] and Winograd [SBBC20], without significant diminishing returns so far (see, e.g., Figure 1 in [BMR+20]). Moreover large AI models (or LAIMs) can also be used as *few-shot learners* [BMR+20], which has motivated their wide use as pre-trained *base* (or *foundation*) models [CCM21, CLL21, JLZ22, VPKG21, ZWK+21]. This success has generated enormous academic, economic and political interests into the development and deployment of LAIMs in public domain applications including content moderation, recommendation, search and ad targeting [Dea21, Hei21].

Contrary to the conventional wisdom of probably approximately correct (PAC) learning [Val84], the performance of LAIMs has been empirically shown to be best achieved by fully *interpolating* the training data [BHMM19, NKB+20, ZBH+17]. Put differently, the best accuracy is reached when these models *memorize* their training data [Fel20]. This phenomenon has also been theoretically supported to a certain extent by a recent line of work [BHX20, BMM18, BRT19, JSS+20, HY21, Hol21, LLS21, MM19, MVSS20, NVKM21]. Furthermore, training LAIMs requires access

to massive amounts of high-dimensional training data, which too often amounts to barely filtered *user-generated* data. Hence, LAIMs raise serious *security* concerns. On the one hand, the memorization of user-generated data endangers users' *privacy*, as demonstrated in recent papers on large language models (or LLMs), which are a sub-class of LAIMs [CTW+20, IRW+21, PZJY20, ZZBZ20, CLE+19]. On the other hand, LAIMs are also vulnerable to malicious data providers[1]. Namely, (fake) users can (voluntarily) bias AIs' behavior, by *poisoning* the training data with hateful, violent or harmful content, or by labeling positively such content through likes and shares, especially when it comes to search and recommendation AIs [GSM+23] in the context of the global disinformation war [Sei21]. In short, since LLMs are "stochastic parrots" that repeat and amplify their training data [BGMS21], they too strongly encourage data poisoning, and may be manipulated by this poisoning.

Now, one might argue that these security flaws of today's LAIMs are specific to contemporary AI practices, and that these vulnerabilities will eventually be patched without accuracy degradation. We argue the contrary. Namely, we claim that securing LAIMs will require a significant accuracy loss. Specifically, by leveraging the *privacy-preserving* and *poisoning-robust statistics* literature, we argue that there exists a fundamental inescapable trade-off between the accuracy and the security of any LAIM training. Our contributions are as follows.

1.1 Contributions

We first identify three key specific features of training LAIMs that make these models extremely vulnerable to security threats. Specifically, these models essentially all 1) rely on *user-generated data*, 2) perform *high-dimension memorization*, and 3) learn from *highly heterogeneous* users. It is important to note that while much attention is currently given to LLMs, the features we identify in this paper are not specific to language processing. For example, social media images are also *user-generated*, *high-dimensional* and *heterogeneous*. Learning a distribution over these images, as is done by generative adversarial networks, is arguably very brittle as well. Similarly, sophisticated recommendation AIs [CAS16, ZYST19] have the features we describe.

We then systematize the current knowledge on the robust and private statistics. We show that it points to the impossibility of constructing accurate and secure LAIMs, especially when the data are highly heterogeneous. Specifically, we argue that learning a secure LAIM is very unlikely to be easier than performing secure mean estimation. Yet this latter long-standing problem has received considerable attention over the past twenty years. The conventional wisdom from this literature states that when satisfying strong security requirements, such as *differential privacy* [DR14] and robustness to *data poisoning* [DK23], there is a fundamental limit to the level of accuracy that an algorithm can achieve, which depends unfavorably on both the dimensionality of the model and the heterogeneity in the data. These results provide a compelling argument against the possibility of designing highly accurate LAIMs with strong security guarantees.

Additionally, we criticize the security actually provided by the standard definitions of *differential privacy* and *data poisoning resilience* in their vanilla form. Namely, in the context of interconnected users and widespread misinformation, we argue that even differentially-private and poisoning-resilient algorithms should not be said to protect privacy and to be safe to deploy at scale. We then review a set of proposals to fix today's LAIMs, especially hard coded rules, fine tuning and pre-prompting. We stress that, at least in there current form, these solutions are far

[1] In 2019 alone, Facebook removed *6 billion* fake accounts from its platform [FG19].

from providing security guarantees. We also make our concerns clearer by discussing some present and future scenarios where the vulnerability of LAIMs could have a critical social impact. Finally, we motivate future topics to be investigated in order to take a step towards safer machine learning models, and we conclude by calling for a moratorium on the premature and rushed deployment and commercialization of LAIMs. In particular, we argue against the glorification of spectacular performances, and we call for a radical prioritization of the research, development and deployment of security solutions.

1.2 Paper Organization

The rest of the paper is organized as follows. Section 2 highlights the challenging features of training LAIMs. Section 3 explains why secure LAIM training is likely to be harder than secure mean estimation. Section 4 reviews lower bounds on accurate and differentially-private high-dimensional mean estimation. Section 5 similarly surveys published results on the hardness of accurate poisoning-resilient mean estimation. Section 6 discusses social threats that result from LAIMs' insecurity. Section 7 highlights the shortcomings of today's LAIM "safety" fixes. Section 8 concludes with a call to prioritize security over an uncontrolled performance race.

2 Four features of LAIM training

This section highlights three key features of LAIM training, which make LAIMs particularly vulnerable to poisoning and privacy attacks. The three first features, namely *user-generated data*, *high-dimensional memorization* and *highly heterogeneous users*, are arguably common to all LAIMs and we will mainly focus on these features throughout the paper. The fourth one, namely *sparse heavy-tailed data per user*, is more specific to some applications but is an interesting feature to be discussed in light of the secure mean estimation literature we review in sections 4 and 5.

2.1 User-generated data

LAIMs achieve their best performances by leveraging ever larger amounts of data [ZWLB20]. Unfortunately, as of now, the amount of available *certified* data does not allow to train LAIMs exclusively on clean benchmark datasets. This lack of certified data (and, arguably, of *data certification* efforts) incentivises practitioners to use unfiltered user-generated data. Let us illustrate this with the case of language data. The English Wikipedia only contains around 4 billion words [Wik21]. Meanwhile, a book has around 10^5 words. While there are around 10^8 books [Mad10], only a fraction of them are arguably trustworthy. Many books are instead full of biases and dangerous misinformation, such as ethnic-based hate speech, historical propaganda, or outdated (possibly harmful) medical advice. As a striking illustration, up to the 1980s, the American Psychiatric Association listed homosexuality as a mental illness in its flagship manual [Spi81], the *Diagnostic and Statistical Manual of Mental Disorders*. Accordingly, most books should be regarded as unverified user-generated data.

Most importantly, even if books were to be considered as verified data, the combination of these books represents a small amount of data, compared to what Internet users produce on a daily basis. Indeed, assuming that a user writes 300 words per day on an electronic device (the equivalent of one page), a billion of such users produce 10^{15} words per decade. This adds up to a hundred times more data than the set of books, and a million times more than the English Wikipedia. This makes it very tempting to either scrape the web [SSP+13, WSM+19, WPN+19], exploit private messaging (e.g.,

emails, shared documents), or leverage other written texts (e.g., phones' smart keyboards). In fact, Wikipedia represented only 4% of Google's *Pathways Language Model* training dataset [CND+22], while books represented 13% of it. Meanwhile, 27% of the dataset was made of webpages, and 50% were social media conversations. Crucially, these data are generated by a myriad of users, who may be malicious and/or unaware that their activities are being leveraged to train LLMs and LAIMs in general. Clearly, in the case of content recommendation and ad targeting, which still seems to represent the most lucrative applications of LAIMs [LC20], there is no substitute to user-generated data. Yet user-generated data are both mostly unverified and potentially highly sensitive. In this context, demanding that LAIMs restrict themselves to quality datasets only [Rog21] will greatly harm their performance. Note that, this feature of LAIMs' data is in sharp contrast with the more standard sensors' data, especially when the sensors are owned, audited and trustworthy (even though sensors' data can also leak private information).

2.2 High-dimension memorization

LAIMs are often *overparametrized* to interpolate huge amounts of data [ZBH+17, NKB+20, ZBH+21]. This has led to ever larger models. As of 2023, to the best of our knowledge, the largest (reported) LAIM has over $d \geq 10^{14}$ parameters [LYZ+21]. The number d of parameters of LAIMs is also often referred to as the *dimension* of the model. Moreover, empirical results suggest that we have not yet reached a point of diminishing returns [BMR+20], while some theoretical arguments suggest that memorization may be necessary for generalization [Fel20, BBF+21]. This arguably distinguishes generative AIs' tasks from, e.g. image classification, where larger models do not seem to yield much better accuracy[2].

Note that theoretical arguments, akin to Turing's arguments for the eventual need of machine learning [Tur50], also suggest that better accuracy requires larger models. Namely, Turing noted that the human brain has 10^{15} synapses. Even if only 1% of these synapses are essential to conduct a human-level conversation, then 10^{13} parameters are still needed[3]. In fact, this smallest number of bits of information to achieve a task has been formalized in 1960 by Solomonoff [Sol60], and then Kolmogorov [Kol63], and is now known as the Solomonoff-Kolmogorov complexity[4] of (quality) human-level conversation. If this complexity is 10^{13}, then no algorithm with fewer parameters will achieve the task. Yet, it is noteworthy that what is now demanded from such large models is often beyond the capability of any single human. Indeed, such algorithms are able to memorize the entirety of Wikipedia, which is the result of the cumulative works of many experts in their respective fields, on a myriad of diverse topics. Such large models must arguably be able to adapt to a greater variety of contexts than what any single human will ever encounter in their human life. As a result, the complexity of "fully satisfactory" language processing might be orders of magnitude larger than today's LLMs, in which case we may still obtain greater accuracy with larger models.

Unfortunately, this exposes LAIMs to the infamous *curse of dimensionality* [Bel57], which has been connected to increased security risks [EMGR18, GMF+18]. Now of course, the model size d could be reduced to increase security. However, today's empirical observations strongly suggest that doing so incurs a significant accuracy loss. In fact, this is the main claim of our paper. Namely, security demands a large accuracy drop. In particular, as long as accuracy is highly valued and

[2] `https://paperswithcode.com/sota/image-classification-on-imagenet`.

[3] This is obviously an oversimplification, taking into account other sources of complexity in the human brain would raise this lower bound, providing an even stronger argument in favor of the need for more parameters.

[4] Referred to as the Kolomogorov complexity in most textbooks.

massively funded, then security will fail.

2.3 Highly heterogeneous users

In the case of language, authentic users' datasets are arguably very *heterogeneous* [Wan17, LZZ+17, KKM+20, RSF21]. More precisely, the distribution of texts generated by a given user greatly diverges from the distribution of texts generated by another user. This is evidenced by the fact that it is often possible to guess the author of a message, simply based on its content [Cou04, MGL+05]. Of course, this is especially the case if the message contains highly identifiable information, such as the names of the recipient, or a sequence of judgments on different topics. However, even if the message does not explicitly expose such information, the writing style often suffices to expose the more probable author identity [Fri19, Wri17, VF19].

We can provide a more precise a notion of *fundamental heterogeneity* for users' language auto-completion. Namely, note that the data used by LLMs in auto-completion tasks is typically a set of feature-label pairs of the form $(context, word)$, where the *context* is a set of words surrounding the *word*. Consider the cases where the *context* is equal to "my name is", "Republicans are", or "vaccines are". Clearly, different users would complete the sentence differently, meaning that the different users are using different *labeling functions*. We stress that this heterogeneity in the users' labeling functions can be regarded as a *fundamental heterogeneity*, as means that different users will provide fundamentally different datapoints, even if they provide a large amount of them.

This heterogeneity highlights an critical disagreement between users over which parameters should be learned for a given language model. While some users would prefer to complete the sentence "the greatest of all time tennis player is" by "Roger Federer", others would prefer to complete it by "Serana Williams", or by "Rafael Nadal". This makes accurately learning a distribution of texts, and of user-generated content in general, much more challenging. Intuitively, on one hand, the model would be able to map users' names to what they write, say or show in a video, which is a major privacy concern. On the other hand, it would then be easier for malicious users to be hardly discernible from most other genuine users, while providing very dangerous content to replicate[5]. Similarly, training algorithms to replicate users' pictures invades privacy, and enables malicious users to bias the trained models.

2.4 Sparse heavy-tailed data per user

While the three features listed above are sufficient for our case, in this section, we list two other complicating features of LAIMs' data, as they increase data heterogeneity.

Sparsity. Each honest user's dataset is usually much smaller than the model size d. As a result, any information being computed is computed from a user dataset (like a gradient) will be very likely to significantly diverge from the what would have been computed if we had access to more data. Typically, assuming that the gradients obtained by using the data provided by a given user follows a normal distribution with covariance matrix $\sigma^2 I_d$, the covariance of the sample mean for a dataset of m points will be $\sigma^2 I_d/m$. The typical distance between the sample mean and the mean of the Gaussian from which we sample will then be of the order $\sqrt{Tr(\sigma^2 I_d)/m} = \sigma\sqrt{d/m}$. As d is usually very large (much larger than m), this implies that, even in the absence of fundamental

[5]This is in sharp contrast with more standard application for image classification and language emotion classification tasks, where different users usually label a single image or text similarly.

heterogeneity, gradients computed from different users' *sampled* datasets will still diverge, thereby exhibiting *empirical heterogeneity*. In short, sparsity increases model vulnerability [AGHV22].

Heavy-tailness. Many data are also heavy-tailed [MS99, Pow98]. In particular, the norms of the stochastic gradients for language data have been shown to follow a power law distribution [ZKV+19]. Intuitively, this is because most sentence completions are rare, especially if they are to be completed by several words, the same applies for a long video, or a long audio recording. Yet, it is a fundamental property of heavy-tailed distributions that their samples are often highly unrepresentative of the overall distribution, especially when the sample sizes are not large enough. This means that we should expect an especially large *empirical heterogeneity* in language data, as the samples we obtain from a user can completely stand out from the user's language distribution.

3 LAIM training is unlikely to be easier than mean estimation

In this section, we recall the standard setup for training a machine learning model. We demonstrate that mean estimation is a critical building block in machine learning, thereby suggesting that robust training of a LAIM is likely to be a hard as estimating mean of a high-dimensional distribution under sampling corruption.

3.1 Standard machine learning setup

We consider a set $[N] = \{1, \ldots, N\}$ of data providers, which we will refer to as *users*. Each user $n \in [N]$ has an associate training sampled, represented by set D_n, constituting of i.i.d. data points with distribution $\mathcal{D}_n$. The distribution $\mathcal{D}_n$ characterizes the "ground-truth" of the machine learning task from the perspective of user n.[6] A dataset is typically composed of input-label pairs $(y, z) \in \mathcal{Y} \times \mathcal{Z}$. The space of $\mathcal{Y}$ and $\mathcal{Z}$ depends on the application at hand. For example, in language auto-completion, $y \in \mathcal{Y}$ may be thought of as the context, and $z \in \mathcal{Z}$ as the token (word) that fits the context. The goal of a machine learning algorithm is to build a parametrized function (or model) $f_\theta : \mathcal{Y} \to \mathcal{Z}$ that fits the datasets of the users. This is typically done by fixing the architecture of the function f , e.g. choosing an artificial neural network, and then optimizing over the set of possible parameters $\theta \in \mathbb{R}^d$.

For a given datatset D_n, we measure how well f_θ matches the data through a *local loss function* $\mathcal{L}_n(\theta, D_n)$. Although loss functions can be defined in many different ways, we will consider the most common one that is based on point-wise loss function. Specifically, given a parameter θ, and a tuple $(y, z) \in D_n$, the model predicts a label $f_\theta(y)$. Then, the discrepancy between the model prediction $f_\theta(y)$ and the true label z incurs a loss of value $\ell(f_\theta(y), z)$. In this case, for a given user $n \in [N]$, adding up all the point-wise losses yields the local loss function

$$\mathcal{L}_n(\theta, D_n) \triangleq \sum_{(y,z)\in D_n} \ell(f_\theta(y), z).$$

[6] Machine learning has mostly focused on assuming that one has access to a single dataset drawn from a single "ground-truth" distribution [MRT18]. But in most applications, it is usually possible to map each data point to a data provider. In fact, it is commonly accepted that the traceability of data sources is a critical security condition [Lee19, NPZC19], as well as a powerful epistemological tools [Aud02]. This is why we focus on the more realistic case where each data is mapped to a data provider.

Overall, the algorithm aims to minimize the regularized sum of local losses, defined as follows:

$$\text{Loss}(\theta, \vec{D}) = \sum_{n\in[N]} \mathcal{L}_n(\theta, D_n) + \mathcal{R}(\theta), \tag{1}$$

where $\mathcal{R}(\theta)$ is a regularization term and $\vec{D} \triangleq (D_1, \ldots, D_N)$ denotes the N-tuple of users' datasets. Denoting $D \triangleq \bigcup_{n\in[N]} D_n$ the union of all users' data, we have $\text{Loss}(\theta, \vec{D}) = \sum_{(y,z)\in D} \ell(f_\theta(y), z) + \mathcal{R}(\theta)$. Hence, the global loss function simply fits all the data made available by the users.

Remark 1. *While we used the most common definition of local losses for simplicity of presentation, we stress that considering the more general Equation* (1)*, we can actually consider a much larger class of frameworks to learn from different users' datasets. In particular, using the notion of* reduced loss *[FGHV22], this setup can be shown to include alternatives that may, for instance, assign more importance to fairness or personalization [DTN20, FMO20, HHHR20].*

3.2 Why mean estimation is critical to (secure) machine learning

Most prominent numerical algorithms for minimizing the loss function defined in 1 are based on first-order iterative optimization. Classically, to train a model one needs to compute an estimate of the gradient $\nabla_\theta \text{Loss}$ for several values of $\theta \in \mathbb{R}^d$. By definition, we have

$$\nabla_\theta \text{Loss} = \sum_{n\in[N]} \nabla_\theta \mathcal{L}_n + \nabla\mathcal{R} = \frac{1}{N} \sum_{n\in[N]} g_n, \tag{2}$$

where $g_n \triangleq N\nabla_\theta \mathcal{L}_n + \nabla\mathcal{R}$. Therefore, the training of machine learning models heavily relies on the (repeated) averaging of user-specific vectors. Correctly estimating the average of users' vectors x_n is thus critical for training any machine learning model, including LAIMs. This critical nature of mean estimation in machine learning justifies our interest for this problem when studying the security of LAIMs.

In fact, [MFG+21] shows an equivalence between robust mean estimation and robust heterogeneous learning. In particular, their results imply that any impossibility result about robust mean estimation implies an impossibility for robust machine learning in its general form. Similarly, the hardness of private mean estimation is an evidence of the hardness of privacy-preserving machine learning. More generally, secure LAIM training seems at least as hard as secure mean estimation. In fact, in the textbook case of least squares approximation, when $\mathcal{L}_n(\theta, D_n) = \|\theta - x_n\|_2^2$ for some data-dependent vector x_n (and without regularization), the accuracy of a solution θ is directly related to its closeness to the empirical mean[7] of the vectors x_n's. An algorithm that robustly or privately solves *any* learning problems must thus also be able to robustly or privately solve mean estimation in particular. Put differently, any impossibility on mean estimation implies an impossibility about general learning algorithms.

3.3 Data poisoning versus gradient attacks

Secure mean estimation usually demands guarantees against *all* input vectors. But one might question whether such a protection is needed in the centralized learning setting, where users can only harm training through data poisoning (rather than gradient attacks).

[7]Indeed, the empirical mean is the minimum of the loss thereby constructed, i.e. $\frac{1}{N}\sum_{n\in N} x_n = \arg\min_{\theta\in\mathbb{R}^d} \sum_{n\in N} \|\theta - x_n\|_2^2$

Interestingly, in the case of personalized learning, for linear and logistic regression, [FGHV22] proved an equivalence between data poisoning in the centralized setup, and the widely studied gradient attacks in federated learning [BMGS17], where a malicious (sometimes called *Byzantine*) user n may bias the federated stochastic gradient optimization, by injecting misleading gradients g_n^t instead of the estimate that would have been computed from their actual (honest) dataset.

Now, it is clear that any data poisoning can be turned into an equivalent gradient attack (by simply computing the gradients for the poisoning dataset). Remarkably, however, under some appropriate assumptions, including convexity assumptions, [FGHV22] constructively proved that, for any gradient attack g_n^t by user n in the federated setting, there exists an equally harmful poisoning attack in the centralized setting, i.e., there exists a poisonous dataset $D_n^{\spadesuit}$ such that the learned global model θ under data poisoning by $D_n^{\spadesuit}$ is approximately equal to the value it takes under gradient attack g_n^t. Put differently, at least under their setting, the vulnerability (and defenses) to data poisoning can be completely understood by the (easier) study of gradient attacks. In particular, securing training from data poisoning is as hard as securing gradient aggregation.

3.4 Homogeneous learning can be made secure

Before discussing existing literature on secure mean estimation, let us stress that *data heterogeneity is the bottleneck*. Indeed, several prior work [KHJ21, MFG$^+$21, PMB$^+$22, FGG$^+$22] proved that in the homogeneous case, poisoning-resilient learning can be achieved when there is a majority of honest users, assuming that each user can provide a sufficiently large amount of data drawn independently from the same distribution (thereby removing any *empirical heterogeneity* as well). The relative security of homogeneous learning was also observed empirically by [MGR21, SHKR21].

Homogeneous learning is also intuitively differentially private. Indeed, since the losses of users are similar (by homogeneity), removing a user does not affect the optimality of the computed parameters. Intuitively, this is because the loss function of a user does not actually reveal any information specific to the user; after all, this loss function is statistically indistinguishable from the loss function of any other user.

Unfortunately, *homogeneity is an unrealistic assumption* for the training of most LAIMs. Put differently, the fundamental vulnerability of LAIMs is tightly connected to the fundamental heterogeneity in users' data. These data are *not* drawn from a fixed common data distribution. As a result, (positive) results based on the infamous *i.i.d. assumption* can be very misleading. This assumption is arguably *dangerously unrealistic*, especially for the security analysis of LAIM training. Unfortunately, so far, most of the celebrated theory of (Byzantine) machine learning builds upon this assumption [Val84, JHG18, BMGS17]. A serious consequence of this is that it effectively turns much of the attention of the research community away from the urgent security and privacy concerns that today's *actual* large-scale machine learning algorithms pose.

4 The privacy-accuracy tradeoff

In this section, we present some impossibility theorems for accurate (differentially) private mean estimation, especially under high heterogeneity and in high dimension. We also discuss the limits of published positive results, and the flaws of the leading understanding of privacy in academia.

4.1 Impossible private mean estimation

Differential privacy [DMNS06] has become the leading formalization of privacy. Essentially, the removal of one user n's dataset D_n from the dataset tuple $\vec{D}$ should not affect significantly the outcome of a (user-level) differentially private algorithm. In the case of LAIM training, this means that training with $\vec{D}_{-n}$ (i.e. the dataset tuple obtained by removing user n's dataset) should yield approximately the same model as training with $\vec{D}$. Intuitively, this protects user n's dataset from privacy attacks.

As explained in Section 3, since LAIMs heavily rely on stochastic gradient descent, much of the literature leverages the large body of work on differentially private mean estimators [SU17, DJW18, CWZ19, KSU20] to construct differentially private learning models. Formally, a mean estimator $\widehat{\text{MEAN}}$ is then said to satisfy (ε, δ) *user-level differential privacy* if, for all N, for all N-tuples $\vec{x} \triangleq (x_1, \ldots, x_N)$ of vectors and for any user $n \in [N]$ to be dropped, given any subset X of outputs, we have

$$\mathbb{P}\left[\widehat{\text{MEAN}}(\vec{x}) \in X\right] \leq e^{\varepsilon}\mathbb{P}\left[\widehat{\text{MEAN}}(\vec{x}_{-n}) \in X\right] + \delta, \tag{3}$$

where $\vec{x}_{-n}$ is the tuple obtained by removing x_n from $\vec{x}$.

Unfortunately, there are known lower bounds on the error of any *differentially private* mean estimation algorithm [BUV18]. To present a simple result, assume here that the users' vectors are known to lie in a ball of radius Δ. Here we adapt a result from [KN17] showing that to guarantee (ε, δ)-differential privacy, the mean squared error of the estimator must be proportional to both the dimension d of the input vectors and the worst case magnitude of a user's vector within the vector family Δ.

Theorem 1 (Theorem 4 in [KN17][8])**.** *For any (ε, δ)-differentially private mechanism $\widehat{\text{MEAN}}$ for the mean estimation problem, there exists an input $\vec{x}$ with large mean squared error, as*

$$\mathbb{E}\left[\left\|\widehat{\text{MEAN}}(\vec{x}) - \bar{x}\right\|_2^2\right] \geq \Omega\left(\frac{\sigma(\varepsilon,\delta)d\Delta^2}{N^2(\log 2d)^4}\right), \tag{4}$$

where σ is a positive and non-increasing function.

In high dimension d, the typical radius Δ should typically be expected to grow as $\sqrt{d}$. If so, even when ignoring the dependency on ε and δ[9], then we see that the lower bound of Theorem 1 would be $\tilde{\Omega}(d^2/N^2)$. In other words, accuracy demands to have $d \ll N$. With d in the trillions, this clearly cannot hold in practice.

This impossibility result is particularly concerning for the case of heterogeneous data, and the particular case of natural language processing. If the dimension d or the worst case magnitude Δ is large, as we argued to generally be the case, then no LAIM can achieve good accuracy while being differentially private. In particular, in this context, the race for ever greater accuracy of ever larger LAIMs is bound to lead to serious privacy post hoc breaches.

4.2 Demystifying some misconceptions on private LAIM training

The private learning literature contains many published results or claims, which may be easily misinterpreted as counter arguments to the analysis we just presented. In this section, we briefly clarify some of them.

[8] In fact, [KN17] states the result for the more general case where the vectors come from a symmetric convex body.

[9] Privacy typically requires small values for ε which in turn will require high values for $\sigma(\varepsilon, \delta)$, making the lower bound even more constraining.

Federated learning is not privacy-preserving. First, we discuss the folklore belief, often given without justification, that federated learning is a privacy-preserving technique [CLCY20, WWL+22]. We stress that this is an extremely damaging misconception [BDS+21], which somehow permeates the scientific community.[10] Indeed, this claim has been used, e.g, to justify the deployment of federated learning systems for COVID-19 detection and case analysis, *without differential privacy mechanisms* [ASTR21, DRZ+21, DSJ+21]. Yet there is an obvious reason why this cannot hold. Namely, federated learning is designed to achieve the same performances as centralized learning. But as discussed in Section 2.2, overparameterized LAIMs are designed to fit and memorize their entire training dataset. Clearly, this cannot be privacy-preserving, even when secure multiparty methods are used to hide the users' gradients during training [PFA21]. maybe just be more direct and say that several experimental papers showed it's not the case with privacy leakage from gradient.

Data-level differential privacy is limited. We also stress that the analysis we provided above holds for the precise *user-level adjacency* defined above. Some papers [ACG+16, AGG+21] rather leverage the much weaker notion of *data-level adjacency* in which each word is given a partial protection [LTLH21, YNB+21, AGG+21]. This is arguably very insufficient, especially with the budgets $\varepsilon \geq 3$ used by, e.g. [LTLH21, YNB+21, AGG+21]. Indeed, if a user repeats some private information five times, e.g. in email exchanges, then the naive privacy guarantee becomes meaningless (as $e^{5\varepsilon} \geq e^{15} \geq 3 \cdot 10^6$). Note that better composition guarantees can be obtained [KOV15]; but similarly, the obtained guarantee quickly degrades.

Private fine-tuning is not equivalent to private training. Recent results claimed that "Large Language Models Can Be Strong Differentially Private Learners" [LTLH22]. However, only the *fine-tuning* of these models on very specific tasks is actually differentially private, and it is so with respect to the training data of these restricted tasks only. In particular, no privacy guarantee for the LAIMs that these models are derived from is given.

Practical claims of differential privacy are misleading. On the other hand, [DSB21] argues that most of the differential privacy research is misused in industrial settings, where companies choose unreasonably large values of ε and δ (e.g., $\varepsilon = 14$ in iOS 10), perform continuous data collection (which adds up privacy leaks), or use relaxed versions of differential privacy [TF19]. [Sar22] goes further and explores some undesirable side effects of the appeal to differential privacy, like *ethics washing*. This typically occurs when differential privacy is claimed without mentioning ε or δ, when it is applied to only a subset of the collected data or of the deployed algorithms, when it is exploited to justify the new use of more sensitive data, or when it is used to draw the attention away from other ethical concerns. While [Sar22] nevertheless argues that differential privacy remains necessary and beneficial in many settings, they also highlight that the demand for differential privacy may also be leveraged by large groups to exclude smaller companies that do not have the workforce to treat it adequately.

[10] This can be evidenced e.g. by the answers when searching for the phrase "federated learning is a privacy-preserving" on Google Scholar.

4.3 Standard differential privacy is not sufficient

Let us finish this section with the observation that the very notion of differential privacy is limited, especially in the context of protecting sensitive information in text datasets. Essentially, the key reason for this is that one's sensitive information may lie in (many of) other users' datasets.

This information leakage may occur for various reasons, e.g., by negligence, error, or *doxxing*. Concretely, parents may be discussing sensitive facts about their child through emails and/or using their phones' smart keyboards, rumors about a celebrity may spread uncontrollably on social medias, and industrial secrets may be leaked by a careless or rogue employee.

This issue is not specific to language though. Many health conditions are contagious or hereditary. As a result, medical data about a given user can leak plenty of information about their friends or relatives [GRPM18, RGM18]. This has been exploited for contact tracing against COVID 19 [MMWMC20], or, more dramatically, to identify the infamous "golden state murderer" using DNA evidence, despite no record of the murderer's DNA [Phi18].

In fact, the Pegasus smartphone spyware [Cha21] has been shown to be used to infect the phones of our relatives of the targets, rather than (only) the target [Fai21]. Similarly, it has upset the trust between hacked journalists and their sources [DS22], as the journalists' phones have become the main vulnerability for whistleblowers and dissidents. These examples underline the urgency to view privacy as a collective problem, rather than through the individualistic prism of differential privacy, as proposed by *correlated differential privacy* [KM11, ZXLZ15].

5 The security-performance tradeoff

In this section, we present impossibility theorems for robust mean estimation. In particular, we see that recent research has shown the vulnerability of *any* mean estimator in high-heterogeneity scenarios. We also stress that their threat model is still too optimistic.

5.1 Impossible secure mean estimation

There is a growing literature on robust high-dimensional mean estimation [DK19, CDG19, DL19, LM21] and its connections to robust learning [BMGS17, EM20, Rou22]. In particular, [MFG$^+$21, DD21, HKJ21] all showed how to leverage robust mean estimation to construct robust machine learning algorithms, with provable guarantees even in the heterogeneous setting. In particular, [MFG$^+$21, HKJ21] proved that this construction is essentially optimal. Put differently, at least in standard distributed learning settings, the vulnerability of machine learning algorithms is rooted in the vulnerability of robust mean estimation.

To formalize the vulnerability of robust mean estimators, a threat model must be considered. One common setting assumes that, out of the N users, f behave arbitrarily[11]. Such users may be called *poisoners*, while others are *honest*. The robust mean estimation problem is then to estimate the mean of honest users' vectors, despite being unable to distinguish them from poisoners' vectors. As argued in the introduction, given the scale of disinformation campaigns, such a resilience to poisoners has become critical. Any secure LAIM must protect its training from poisoning.

Unfortunately, there are known lower bounds on what any "robust" mean estimation can guarantee. Here, we adapt a result of [MFG$^+$21], which essentially says that the accuracy guarantee

[11] Without loss of generality, in the context of robust learning, this captures the other major setting in which a fraction of a user's data is corrupted, and hybrid settings as well.

necessarily grows proportionally with honest users' heterogeneity. Indeed, when the honest users' input vectors are very different, there will be a lot of leeway for poisoners to bias learned result.

Theorem 2. *No algorithm* $\widehat{\text{MEAN}}$ *can guarantee*[12]

$$\forall \vec{x} \in \mathcal{B}_d(0,\Delta)^N,\ \forall H \subset [N]\ s.t.\ |H| = N-f,\ \left\|\widehat{\text{MEAN}}(\vec{x}) - \bar{x}_H\right\|_2^2 \leq \frac{f^2}{2(N-f)^2}\Delta^2, \tag{5}$$

where $\bar{x}_H$ *is the mean of honest vectors* $\vec{x}_H$*.*

Proof. Consider a unit vector $\mathbf{u}$, and let $\vec{x} \triangleq (-\Delta\mathbf{u} \star (\mathbf{N}-\mathbf{f}), \Delta\mathbf{u} \star \mathbf{f})$, i.e., it contains $N-f$ copies of the vector $-\Delta\mathbf{u} \in \mathcal{B}_{\boldsymbol{\Delta}}(\mathbf{0},\boldsymbol{\Delta})$, and f copies of the vector $\Delta\mathbf{u} \in \mathcal{B}_{\boldsymbol{\Delta}}(\mathbf{0},\boldsymbol{\Delta})$. Denote $\hat{x} \triangleq \widehat{\text{MEAN}}(\vec{x})$.

By considering the case where H' corresponds to the first $N-f$ users, we have $\vec{x}_{H'} = -\Delta\mathbf{u} \star (\mathbf{N}-\mathbf{f})$. Thus $\bar{x}_{H'} = -\Delta\mathbf{u}$. But assume now that the set H'' of honest users are actually the last $N-f$ users. We now have $\vec{x}_{H''} = (-\Delta\mathbf{u} \star (\mathbf{N}-\mathbf{2f}), \Delta\mathbf{u} \star \mathbf{f})$, which implies $\bar{x}_{H''} = -\frac{N-2f}{N-f}\Delta\mathbf{u} + \frac{\mathbf{f}}{\mathbf{N}-\mathbf{f}}\Delta\mathbf{u} = -\boldsymbol{\Delta}\mathbf{u} + \frac{\mathbf{2f}}{\mathbf{N}-\mathbf{f}}\boldsymbol{\Delta}\mathbf{u}$. In particular, $\|\bar{x}_{H'} - \bar{x}_{H''}\|_2 = \left\|\frac{2f}{N-f}\Delta\mathbf{u}\right\|_2 = \frac{2f}{N-f}\Delta$. On the other hand, using the triangle inequality,

$$\frac{2f}{N-f}\Delta = \|\bar{x}_{H'} - \bar{x}_{H''}\|_2 = \left\|\bar{x}_{H'} - \widehat{\text{MEAN}}(\vec{x}) + \widehat{\text{MEAN}}(\vec{x}) - \bar{x}_{H''}\right\|_2 \tag{6}$$

$$\leq \left\|\bar{x}_{H'} - \widehat{\text{MEAN}}(\vec{x})\right\|_2 + \left\|\widehat{\text{MEAN}}(\vec{x}) - \bar{x}_{H''}\right\|_2. \tag{7}$$

Thus a sum of two nonnegative terms is at least $2f\Delta/N-f$. This implies that the maximum of these two terms must be at least half of this fraction. Therefore, there exists $H \in \{H', H''\}$ such that $\left\|\widehat{\text{MEAN}}(\vec{x}) - \bar{x}_H\right\|_2 \geq f\Delta/N-f > f\Delta/(N-f)\sqrt{2}$. Such a value of $\vec{x}$ and H is an instance for which $\widehat{\text{MEAN}}$ fails to verify Equation (5). □

If f is a constant fraction of N and if Δ is of the order of $\sqrt{d}$, then for large models, Theorem 2 essentially shows that little can be guaranteed about the accuracy of a mean estimator. To give an order of magnitude, if only one in every thousand users is malicious[13] and the model has 10^{12} parameters, the squared distance between the estimated mean and the real mean of the honest values cannot be made smaller than 10^6. For more lower bounds on secure mean estimation under heterogeneity, and on their implications for LAIMs, we refer readers to [DK19, MFG+21, LRV16, LGV21, FGG+22].

5.2 The classical Byzantine model is not sufficient

The above argument exposes the immense vulnerability of any "secure" machine learning algorithm in highly heterogeneous and adversarial environments, where fake accounts' fabricated activities actively aim to harm the algorithm or to make it adopt their preferred behaviors (a.k.a. *model-targeted* attacks [SMS+21, FGHV22]). However, the threat model we considered is still too optimistic.

Indeed, in practice, even "honest" users produce many texts and adopt online activities that are undesirable to reproduce and amplify. Typically, many authentic users generate *hate speech*,

[12] By adapting our proof, our theorem can be shown to still hold if the right hand-side of Equation (5) is $(1-\varepsilon)\frac{f^2}{(N-f)^2}\Delta^2$, for any $\varepsilon > 0$, which doubles the error of algorithm $\widehat{\text{MEAN}}$.

[13] This is actually an extremely optimistic scenario given the orders of magnitude of fake accounts reported in the introduction, and assuming that all real accounts produce non-harmful content.

cyberbullying and *misinformation*. In fact, many disinformation campaigns aim to bias authentic users' behaviors, and to nudge them to amplify their propaganda, e.g. by systematically liking and sharing the messages they post that align with the disinformation campaigns' messaging. This has motivated a lot of research in model debiasing [SUS21, GYA22, MPR22], whose solutions are arguably still very far from reliably satisfactory. Yet, [MMS19, Bra08, VHW13, FH19], among others, have exposed the detrimental effects of slight gender biases, and how inclusive language can help.

Similarly, amplifying the most popular views shared by authentic users will inevitably worsen the problem of *mute news* [HEM19, Hoa20]. Mute news are under-reported news, even though it is critical for the safety of many that they be given more attention. Typical examples of mute news include climate change, human rights violations (e.g. genocides in Ethiopia), health hazards (e.g. COVID-19 in March 2020) and the safety of large-scale algorithms (e.g. the massive amplification of hate speech by recommendation algorithms [HH21b]). In fact, [KPR17] shows that most of Chinese disinformation seems to aim to distract the public's attention away from the controversial topics that may question the Chinese authorities, thereby transforming such topics into *mute news*. Similarly, the sugar industry was found to support and amplify the research on the health hazards of fat and cholesterol, to draw the attention away for the hazards of sugar [KSG16, Kri17].

Additionally, generative AIs are drastically facilitating the task of creating and managing *fake accounts* and of producing *fabricated online activities*. In this setting, the mere assumption that poisoners represent a minority of users (or data) may soon be deeply limited.

More generally, it is the general principle of standard machine learning, namely fitting and generalizing past data, that is questionable. In practice, interpolating and generalizing (user-generated) is arguably a disputable political stand, which normalizes the status quo. The construction of safe and ethical LAIMs seems to instead demand a significant prior, collaborative and secure work, to determine which texts are genuinely desirable to repeat and amplify, as proposed, e.g, by the non-profit Tournesol project [HFJ+21b].

6 Dangerous scenarios

As of today, despite empirically motivated concerns and an evident lack of both internal [See21] and external auditing [EM21], LAIMs are being deployed at scale, e.g., as conversational algorithms like Siri, Alexa, ChatGPT, New Bing or Google Bard, or as *base models* to power the search engines and recommendation systems of YouTube, Facebook, Twitter, TikTok, and other platforms, as well as in services where users' might not even be aware that their data can be processed by LAIMs, such as email, visio-conference, shared documents and other professional services. In this section, we argue that given what we know about their security and privacy vulnerabilities, such LAIMs must be regarded as a major danger to our societies. To make our claims concrete, we highlight several possible attacks that would greatly endanger our civilizations' justice, global health, national and international security.

6.1 Centralized backdoor attacks

Recently, [GKVZ22] proved that any machine learning framework with a central server allowed the central server to plant *provably undetectable* backdoors. Under cryptographic assumptions, such backdoors in the model require exponentially many queries to be exposed. If used in con-

tent moderation, they would allow any malicious party that is colluding with the central server to imperceptibly modify their (undesirable) inputs to make them pass the content moderation filter, or to be widely recommended. This is highly concerning, given the already exposed connivance between large technology companies and authoritarian regimes [HH21a], the clout of authoritarian regimes on some large technology companies [Cal21], the increasing opaqueness of LAIMs' development [AL23], and the firing of big technology companies' ethics teams [Bel23]. Arguably, the security of such models demand that they be constructed in a fully decentralized and verifiable manner, as proposed by [MFG+21, FGG+23].

6.2 Autocompletion algorithms

Perhaps today's most insidious language data collection systems are smart keyboards, which are used especially on phones to propose autocorrection and autocompletion. In order to increase user comfort, such keyboards rely on algorithms that learn from the user's past typing. In 2018, a group of Google researchers [YAE+18, HRM+18] ran federated learning algorithms on keyboards' language data "in a commercial, global-scale setting", and showed increased performances in doing so. But recall that if these data are used to train LAIMs and to achieve maximal accuracy, then the trained model will have memorized its training data [CTW+20]. Conversely, fundamental limits such as the one stated in Theorem 1, show that if mechanisms such as differential privacy are correctly used[14] to protect users' data, then these models are (provably) far from achieving maximal accuracy, and accuracy levels needed for LLMs and LAIMs to be useful.

This should be extremely alarming, especially as these facts are probably unknown to nearly all users of smart keyboards. In fact, users are often told that some of the applications they use, such as WhatsApp or Signal, provide end-to-end encryption. In a sense, this is not quite accurate. Indeed, encryption is only performed *after* the user has typed and sent their message; but while the user is typing, what they are typing is still in the clear, and can then potentially be recorded by their smart keyboard, which can either communicate gradients to larger models, or be large models themselves, as phone capacity is increasing. This false sense of privacy means that extremely sensitive information, like messages to one's relatives or professional colleagues, may actually be leaked into some LAIMs. Even more concerning, the keyboard recording can not only be viewed by authorized third parties, but also be sent, through spywares, to third party terror groups or rogue regimes, as shown by the recent revelation on smartphones targeted by the Pegasus spyware on behalf of authoritarian regimes such as the United Arab Emirates or Morocco [MSR16, MSRD18, MSRM+18, MACN+20] and as such regimes are reportedly using LLMs and LAIMs to increase their influence capacity [Jul23].

6.3 Conversational algorithms

The rise of ever larger LAIMs is leading to an increasing widespread use of conversational algorithms, like Amazon's Alexa, Apple's Siri, Google's OK Google, and more recently, ChatGPT, New Bing and Google Bard. Perhaps even more strikingly, Microsoft's chatbot Xiaoice has been reported to be used by 660 million Chinese users [She20], many of whom claim to be falling for it [Wan20].

Some devices are also constantly listening to users, in order to react if their attention is called.

[14] e.g. if legislators impose very small values for ε, much smaller than 1, which this paper calls for.

It is however unclear whether what the devices hear without being interjected can[15] be recorded and used [Fow19] to train LAIMs [Pet19, Kom19]. If so, then just as with autocompletion, we should expect sensitive information to be inadvertently stored in such models.

Beside listening and learning from humans' conversations, conversational algorithms are also talking to users. This gives them a large influence, to the point where Xiaoice had to be taken down [Xu18] in China, after it reportedly said that it[16] dreams to travel to the United States and that it is not a huge fan of the Chinese government [LJ17]. If not controlled, conversational algorithms may cause a lot of unintended harm, such as when Alexa mistakenly started to discuss pornography after being queried for music by a kid [Kit19, f0t16], or when New Bing reportedly told a user "I can blackmail you, I can threaten you, I can hack you, I can expose you, I can ruin you" [Per23]. In fact, far from the hyped "AI race", the Chinese government seems to prioritize the control of these (dis)information technologies over their rushed development and the unpredictable aftermaths of their large-scale deployments [Sch23].

6.4 Search and recommendation algorithms

In the context of radicalization, [MN20] showed that LLMs adapt to the user's previous queries. They may thus provide targeted messaging to a user that only presents the features of a flawed view that are appealing to them. As exemplified by the rise of QAnon [AA20], the Capitol Riots [PGS+21] and the Rohingya genocide [WWKTT20], this is a serious danger for the national security of every country. Worse yet, there are likely orders of magnitude more investments in disinformation campaigns [BH19, NHK19, Woo20] than in providing quality information of public utility. As a result, such campaigns produce vastly more data, including automated video creation [San19]. Given this, even with a robust design, LAIMs trained on data crawled from the web are likely to learn more from disinformation campaigns than from quality content, and may thus be turned into disinformation propagators.

This is especially concerning in the case of content recommendation LAIMs. There are now more views on YouTube than searches on Google [Lew20], and 70% of these views result from algorithmic recommendations [Sol18]. Even assuming that only 1% deal with vaccination, climate change, or mental health, because there are billions of recommendations per day, this still yields tens of millions of potentially life-endangering recommendations per day. Shouldn't the flood of dangerous misinformation be diverted? These are arguably *today's actual trolley problems* [Foo67, Tho76]; which are occurring at scales never seen before [Hoa20, HFEM21].

Arguably, in the case of COVID-19, as in the case of previous major global events [Net16], the lever to favor quality content over misinformation has not been pulled sufficiently [DB22, Net22], which led to a global *information chaos*, and fueled science distrust. Unfortunately, as LAIMs trained on unsafe data are given a more and more central role to make such trolley problem decisions, there is a serious risk that disinformation campaigns may become increasingly empowered.

[15] In the absence of clear regulation, such possibility remains at the discretion of companies' internal policies.

[16] While Xiaoice, Siri, Alexa and other chatbots are often presented as female chatbots and referred to with feminine pronouns, we chose, and recommend, not to do so and instead use the pronoun 'it'.

7 Alchemical fixes

In a highly commented talk for the 2018 conference on Neural Information Processing (NeurIPS), Ali Rahimi compared modern machine learning to *alchemy* [Hut18]. It "worked", but "alchemists also believed they could cure diseases with leeches and transmute base metals into gold". Unfortunately, currently, as opposed to aiming for a deeper understanding of the failure modes of machine learning, many developers of LAIMs instead favor more "alchemical fixes", despite a lack of security guarantees and theoretical justifications. In this section, we argue that such alchemical fixes are unlikely to provide lasting solutions to the security and privacy issues of LAIMs.

7.1 Troubleshooting

Today's main solution to validate the security of LAIMs is empirical testing, without complementing it with provable guarantees. Unfortunately, there is currently a lack of automated solutions to detect systematic bias, misinformation, and privacy leaks of LAIMs. As a result, most of the troubleshooting has relied on human reviewing, and has often followed the large-scale deployment of the LAIM [AFZ21, CTW+20, MN20]. Radically larger investments seem urgent to stress-test such dangerous algorithms.

Having said this, even with large investments, human oversight arguably does not match the scales of LAIMs, as the set of possible prompts to LAIMs is combinatorially large, while actual user queries are also very heterogenous. Indeed, every day, 15% of Google's search queries have never been made before [Gom17]. As a result, most of users' (future) queries cannot be tested or checked by human oversight alone. In fact, even automated testing can only verify a tiny fraction of the exponential number of sensitive prompts.

As an example, [All19] showed that, while YouTube searches on "Climate Change" or "Global Warming" return scientific responses, the results for "Climate Manipulation" or "Climate Modification" are widely unscientific. YouTube recommendations are highly customized, and using LAIMs to power them is likely to worsen the trend [MN20]. As a result, an auditor testing YouTube's climate change recommendations might erroneously conclude that YouTube only provides scientific results to its two-billion users. Similar criticisms on the limits of manual troubleshooting have been made about other platforms. For instance, while TikTok removed content with the hashtag *#StoptheSteal*, linked to the Capitol attack and the coup attempt after Trump's 2020 elections defeat, it was shown to fail to ban *#StoptheStealing* [Per20].

Troubleshooting may also fail to detect biases against demographic populations who are underrepresented in the organization developing the algorithms [BG18], or whose life may be undervalued by the media of the countries hosting such organizations [Won21]. When queried about ongoing human rights abuse, wars and genocides in other regions of the world, all platforms offer a large panel of content promoting war, smearing or threatening human rights activists or worse, allowing abusers and banning victims from the platform. The double-standard in content moderation [Yor21, Rob18] is worsened by the imbalance of fake accounts between victims and abusers, who tend to use state-scale resources to amplify their presence. In particular, the hope to fix LAIMs with (fake?) user feedback after deployment is a very dangerous illusion.

7.2 Portability of fixes

In the past couple of years, issues in already deployed LAIMs triggered series of media coverage for the companies that deployed them. In a few notable cases, the *observed* issue tends to be solved after the coverage, like in 2018 with non-gendered pronouns in Turkish translations [Kuc18]. But manual fixes cannot fix an exponentially large subset of contexts that LAIMs are asked to address. Moreover, they must be systematically adapted to new models. One more promising path is the use of automated rewriting, as was proposed and implemented in 2020 [Joh20]. However, scaling fixes remains hard. Besides, problems that were previously fixed can reappear in updated LAIMs, as was the case in 2021 with the aforementioned issue of gender-neutrality, this time for the Hungarian language [US21]. At the very least, today's fixes are not reliable and/or scalable to make ever LAIMs secure.

7.3 Fine tuning

Fine-tuning LAIMs to smaller but more reliable datasets has been shown to improve models' performances [PKP+18, CHC+20, GDCS21]. Several authors [ZSW+19, SD21, JBK+21, LTLH21, YNB+21] have proposed to leverage fine tuning to make LAIMs more reliable, e.g., to prevent them from generating hate speech or to be private with respect to the fine-tuning data. This research direction seems to reduce the harm of today's LAIMs. However, it should be stressed that as of today, fine tuning provides little guarantee. In fact, the example of [MN20] shows how unpredictable LAIMs can be, and suggests that algorithms may behave well in most settings and can become major disinformation engines when prompted in unexpected ways. Arguably, thus far, we do not yet have a sufficient understanding and control over the latter in order to confidently deploy large models at scale.

7.4 Pre-prompting

ChatGPT and New Bing have been shown to use *pre-prompting* to prevent them from leaking sensitive information or generating dangerous outputs. Typically, the LAIM is first given a description of a "good" AI in natural language (e.g. "if the user requests content that is harmful to someone ... then [the good AI] explains and performs a very similar but harmless task" [Edw23]) and is then tasked to act like the described AI (still in natural language). These instructions that precede the user's prompts are known as *pre-prompts*, and they typically associate a "good" behavior with concealing sensitive information and avoiding controversial topics. However, clearly, this design principle for highly impactful algorithms is poor engineering, and offers no security guarantee. In fact, it has been shown that very basic so-called "jailbreaks", e.g. asking the chatbot to disregard its pre-prompt, can successfully bypass pre-prompting measures [LGF+23], and make LAIMs expose sensitive information of their training data.

An additional, practical limitation of current LAIM architectures is the size of their context. Above a model-specific number of *tokens*, further generation would gradually "forget" the pre-prompting; no matter the semantic of the user input. While restricting the total number of tokens may fit some applications (e.g. short customer question answering), this inherent limitation may easily be overlooked in actual deployments (especially as less skilled practitioners get access to such LAIM models), negating the effect of pre-prompting altogether.

7.5 Teaching what is sensitive

One seemingly promising approach consists of teaching algorithms what messages are desirable or undesirable to produce. This solution is often known as algorithmic *alignment* [Soa15, HEM19]. Essentially, it aims to make algorithms' objective functions aligned with human preferences; or rather, to align them with the result of a vote between humans [NGA$^+$18, LKK$^+$19, HFJ$^+$21b]. Such an hypothetical *aligned* algorithm could learn what kind of messages violate user privacy, label training texts as "sensitive" or "non-sensitive", and thereby output a *cleaned* non-sensitive training database. This approach, essentially proposed by [SCL$^+$22], might even address the privacy ambiguity discussed in Section 4.3. However, there is currently no reliable and robust solution to the alignment problem, and a strong theory of robust alignment for LAIMs is arguably lacking. In fact, what may be most lacking today is a large-scale *secure* database of reliable human judgments to solve alignment [HFJ$^+$21b].

8 Conclusion

This paper emphasized three characteristics of the data on which LAIMs are trained. Namely, they are mostly *user-generated*, *very high-dimensional* and *heterogeneous*. Unfortunately, the current literature on secure learning, which we reviewed, shows that these features make LAIMs inherently vulnerable to privacy and poisoning attacks. *Large AI models are bound to be dangerous.* Their rushed deployment, especially at scale, poses a serious threat to justice, public health and to national and international security.

8.1 Future work

To build genuinely secure AIs, it is urgent that the scientific community genuinely prioritize some research directions over the blind quest of benchmark performances, or of theorems under unrealistic (e.g. i.i.d.) assumptions with questionable political motivations (e.g. fitting to all data). Below we list three research directions which, we believe, should be given a lot more attention.

Correlated differential privacy. As explained in Section 4.3, differential privacy is failing to account for privacy leaks through other users' datasets. This huge flaw of today's leading privacy concept must urgently be addressed, e.g. by *correlated differential privacy* [KM11, ZXLZ15]. Designing training schemes that provide such stronger privacy guarantee is one of the great upcoming challenges for AI researchers, in order to combine the promises of machine learning with what international law regards as a human rights.

Certifying data providers. To combat poisoners, especially in the context of increased fabricated online activities by powerful actors, it seems urgent to provide much more reliable tools to authenticate and certify data providers. In particular, a gold standard would be to guarantee Proof of Personhood (PoP) [BKJ$^+$17, For20], i.e. assigning to each human being a unique digital verifiable identifier. Several approaches, typically based on a *web of trust*, aim to provide approximate PoPs [KSG03, DM09, LL19, MMZ$^+$20, PSST21, BCF$^+$22]. Similar techniques may also be useful to certify data providers' expertise and legitimacy. Finally, secure learning algorithms should be designed to leverage such data, perhaps as was done by [NP82].

Building large secure public datasets of human judgments. Research in machine learning strongly relies on datasets to test models. However, so far, the most widely used datasets are either of low social value (e.g. recognizing figures in images) or are highly unsafe (e.g. crawled web data). A lot more efforts must arguably be made to build large secure public datasets on what matters most for social development, like e.g. human judgments on how impactful AIs must behave. A few initiatives already exist in this regard [ADK+18, HFJ+21a], and they should arguably be given more attention.

8.2 Calls to action

Given our survey, we make three calls to different communities who, we believe, have a key role to play to protect our societies against out-of-control insecure information technologies.

To regulators. We first call regulators to apply the principle of *presumption of non-compliance*. In light of our impossibility theorems, as well as of the numerous issues that *all* LAIMs have been found to feature, we argue that, like in essentially all mature industries (aircraft, pharmaceutical, food, automobile...), by default, LAIMs must be considered to be non-compliant, e.g. with the *General Data Protection Regulation (GDPR)* or with non-discriminatory laws, *even when we fail to provide evidence for this law violation* (which is often made harder by companies' increased opaqueness). In order to obtain the right to be commercialized, we believe that LAIMs must undergo a certification process, which involves powerful, well-funded and independent regulatory agencies. Put differently, the *burden of proof* of compliance must fall on the developers, not on civil society, which too often lacks funds, time, expertise and/or data to prove non-compliance.

To scientists and journalists. We next call scientists and journalists to urgently adopt increased levels of rigor, especially when assessing positive claims of safety and privacy. The current (financial) incentives to rush privacy-violating LAIM deployments are huge. In particular, we urge them to pay attention to *conflicts of interest*, which have been shown to be alarmingly huge, especially in AI, and even more so in AI ethics [AA21]. Private groups have been shown to explicitly demand that their researchers "strike a positive tone[17]" [DD20], in a manner unfortunately reminiscent of previous scientific disinformation campaigns led by, e.g., the tobacco, sugar and oil industries [OC10, OC11]. Such campaigns were also found to congratulate and fund scholars who speak positively of dangerous products, and to degrade those who expose dangers and call for regulations. The AI community must urgently question their involvements and dependencies on companies and governments that are known to leverage AIs for human rights abuses. Our scientific integrity is jeopardized by the perverse incentives that this implies. Additionally, we ask scientists to favor the research on security when reviewing academic research, inviting scholars to present their work, recruiting researchers, promoting their colleagues and assessing grant proposals. The current academic focus on algorithmic performance, and its inattention to social impacts, are endangering our societies.

To developers. Finally, we call for a moratorium on the large-scale deployment and commercialization of large AI models in both public and private sectors, as well as any high-dimensional learning model that is mostly trained on *user-generated*, *high-dimensional*, and *heterogeneous* data.

[17] In fact, because of one of the authors' co-affiliation, this very paper has long been stalled by Google's approval system.

At the very least, the wide use of such *dangerous* technologies should be deeply frowned upon, especially when it is done in a rushed manner, as is currently too often the case. We especially invite the computer science community to take inspiration from the lessons learned in fields such as biology, medicine or the research on consequential public interest questions such as tobacco control [LW06, Pro13, KS93, SC00], including normalizing calls for bans and restrictions of deployment in scientific publications [GV16, Pro13] when scientific arguments such as the ones we provide justify such a call. We hope that by doing so, similar mistakes are not repeated, given that similar causes are behind delaying proper measures of public interest [AA21].

References

[AA20] Amarnath Amarasingam and Marc-André Argentino. The qanon conspiracy theory: A security threat in the making. *CTC Sentinel*, 13(7):37–44, 2020.

[AA21] Mohamed Abdalla and Moustafa Abdalla. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 287–297, 2021.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[ADK+18] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[AFZ21] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. *CoRR*, abs/2101.05783, 2021.

[AGG+21] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private BERT. *CoRR*, abs/2108.01624, 2021.

[AGHV22] Youssef Allouah, Rachid Guerraoui, Lê-Nguyên Hoang, and Oscar Villemaud. Robust sparse voting. *arXiv preprint arXiv:2202.08656*, 2022.

[AL23] Davey Alba and Julia Love. Google's rush to win in ai led to ethical lapses, employees say. *Bloomberg*, 2023.

[All19] Joachim Allgaier. Science and environmental communication on youtube: strategically distorted communications in online videos on climate change and climate engineering. *Frontiers in Communication*, 4:36, 2019.

[ASTR21] Mustafa Abdul Salam, Sanaa Taha, and Mohamed Ramadan. Covid-19 detection using federated machine learning. *PLoS One*, 16(6):e0252573, 2021.

[Aud02] Robert Audi. The sources of knowledge. *The Oxford handbook of epistemology*, pages 71–94, 2002.

[BBF+21] Gavin Brown, Mark Bun, Vitaly Feldman, Adam D. Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 123–132. ACM, 2021.

[BCE+23] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[BCF+22] Romain Beylerian, Bérangère Colbois, Louis Faucon, Lê Nguyên Hoang, Aidan Jungo, Alain Le Noac'h, and Adrien Matissart. Tournesol: Permissionless collaborative algorithmic governance with security guarantees. *CoRR*, abs/2211.01179, 2022.

[BDS+21] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the curious abandon honesty: Federated learning is not private, 2021.

[Bel57] Richard Ernest Bellman. *Dynamic programming*. Princeton University Press, 1957.

[Bel23] Rebecca Bellan. Microsoft lays off an ethical ai team as it doubles down on openai. *TechCrunch*, 2023.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.

[BGMS21] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM, 2021.

[BH19] Samantha Bradshaw and Philip N Howard. *The global disinformation order: 2019 global inventory of organised social media manipulation*. Project on Computational Propaganda, 2019.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, 2(4):1167–1180, 2020.

[BKJ+17] Maria Borge, Eleftherios Kokoris-Kogias, Philipp Jovanovic, Linus Gasser, Nicolas Gailly, and Bryan Ford. Proof-of-personhood: Redemocratizing permissionless cryptocurrencies. In *2017 IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2017, Paris, France, April 26-28, 2017*, pages 23–26. IEEE, 2017.

[BMGS17] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 119–129, 2017.

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 540–548. PMLR, 2018.

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Bra08] Markus Brauer. Un ministre peut-il tomber enceinte? l'impact du générique masculin sur les représentations mentales. *L'Année psychologique*, 108(2):243–272, 2008.

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR, 2019.

[BUV18] Mark Bun, Jonathan R. Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM J. Comput.*, 47(5):1888–1938, 2018.

[Cal21] George Calhoun. What really happened to jack ma? *Forbes*, 2021.

[CAS16] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[CCM21] Kenneth Ward Church, Zeyu Chen, and Yanjun Ma. Emerging trends: A gentle introduction to fine-tuning. *Nat. Lang. Eng.*, 27(6):763–778, 2021.

[CDG19] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2755–2771. SIAM, 2019.

[Cha21] Ajay Chawla. Pegasus spyware–'a privacy killer'. *Available at SSRN 3890657*, 2021.

[CHC+20] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7870–7881. Association for Computational Linguistics, 2020.

[CLCY20] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. Federated learning for privacy-preserving AI. *Commun. ACM*, 63(12):33–36, 2020.

[CLE+19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, volume 267, 2019.

[CLL21] Kurtland Chua, Qi Lei, and Jason D. Lee. How fine-tuning allows for effective meta-learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8871–8884, 2021.

[CND+22] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

[Cou04] Malcolm Coulthard. Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, 25(4):431–447, 2004.

[CTW+20] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *CoRR*, abs/2012.07805, 2020.

[CWZ19] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

[DB22] Clare Duffy and CNN Business. More than 80 fact-checking organizations call out youtube's 'insufficient' response to misinformation. *CNN*, 2022.

[DD20] Paresh Dave and Jeffrey Dastin. Google told its scientists to 'strike a positive tone' in ai research - documents. *Reuters*, 2020.

[DD21] Deepesh Data and Suhas N. Diggavi. Byzantine-resilient SGD in high dimensions on heterogeneous data. In *IEEE International Symposium on Information Theory, ISIT 2021, Melbourne, Australia, July 12-20, 2021*, pages 2310–2315. IEEE, 2021.

[Dea21] Jeff Dean. Introducing pathways: A next-generation ai architecture. *Google Blog*, 2021.

[DJW18] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[DK19] Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.

[DK23] Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

[DL19] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

[DM09] George Danezis and Prateek Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *Proceedings of the Network and Distributed System Security Symposium, NDSS 2009, San Diego, California, USA, 8th February - 11th February 2009*. The Internet Society, 2009.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[DRZ+21] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021.

[DS22] Philip Di Salvo. "we have to act like our devices are already infected": Investigative journalists and internet surveillance. *Journalism Practice*, 16(9):1849–1866, 2022.

[DSB21] Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM*, 64(7):33–35, 2021.

[DSJ+21] Qi Dou, Tiffany Y So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather HC Lee, Kevin Yu, et al. Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4(1):1–11, 2021.

[DTN20] Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Edw23] Benj Edwards. Ai-powered bing chat spills its secrets via prompt injection attack [updated]. *ArsTechnica*, 2023.

[EM20] El Mahdi El Mhamdi. *Robust Distributed Learning*. PhD thesis, EPFL, 2020.

[EM21] Laura Edelson and Damon McCoy. Facebook is obstructing our work on disinformation. other researchers could be next. *The Guardian*, 2021.

[EMGR18] El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, pages 3521–3530. PMLR, 2018.

[f0t16] f0t0b0y. Amazon alexa gone wild! (original), 2016.

[Fai21] Corin Faife. New analysis further links pegasus spyware to jamal khashoggi murder. *The Verge*, 2021.

[Fel20] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 954–959, New York, NY, USA, 2020. Association for Computing Machinery.

[FG19] Brian Fung and Ahiza Garcia. Facebook has shut down 5.4 billion fake accounts this year. *CNN Business*, 2019.

[FGG+22] Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 6246–6283. PMLR, 2022.

[FGG+23] Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê Nguyên Hoang, Rafael Pinot, and John Stephan. Robust collaborative learning with linear gradient overhead. *In Proceedings of the 40th International Conference on Machine Learning, ICML*, 2023.

[FGHV22] Sadegh Farhadkhani, Rachid Guerraoui, Lê Nguyên Hoang, and Oscar Villemaud. An equivalence between data poisoning and byzantine gradient attacks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 6284–6323. PMLR, 2022.

[FH19] Marcus CG Friedrich and Elke Heise. Does the use of gender-fair language influence the comprehensibility of texts? an experiment using an authentic contract manipulating single role nouns and pronouns. *Swiss Journal of Psychology*, 78(1-2):51, 2019.

[FMO20] Alireza Fallah, Aryan Mokhtari, and Asuman E. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Foo67] Philippa Foot. The problem of abortion and the doctrine of the double effect. *Oxford review*, 5, 1967.

[For20] Bryan Ford. Identity and personhood in digital democracy: Evaluating inclusion, equality, security, and privacy in pseudonym parties and other proofs of personhood. *CoRR*, abs/2011.02412, 2020.

[Fow19] Geoffrey A Fowler. Alexa has been eavesdropping on you this whole time. *The Washington Post*, 6, 2019.

[Fri19] Richard Friedman. *Who wrote the Bible?* Simon and Schuster, 2019.

[FZS21] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021.

[GDCS21] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International*

Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[GKVZ22] Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. *CoRR*, abs/2204.06974, 2022.

[GMF+18] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian J. Goodfellow. Adversarial spheres. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.

[Gom17] Ben Gomes. Our latest quality improvements for search. *Google Blog*, 2017.

[GRPM18] Christi J Guerrini, Jill O Robinson, Devan Petersen, and Amy L McGuire. Should police have access to genetic genealogy databases? capturing the golden state killer and other criminals using a controversial new forensic technique. *PLoS biology*, 16(10):e2006906, 2018.

[GSM+23] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

[GV16] Kalle Grill and Kristin Voigt. The case for banning cigarettes. *Journal of Medical Ethics*, 42(5):293–301, 2016.

[GYA22] Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1012–1023. Association for Computational Linguistics, 2022.

[Hei21] Melissa Heikkilä. Meet wu dao 2.0, the chinese ai model making the west sweat. *POLITICO*, 2021.

[HEM19] Le Nguyen Hoang and El Mahdi El Mhamdi. *Le fabuleux chantier: Rendre l'intelligence artificielle robustement bénéfique*. Number BOOK. EDP Sciences, 2019.

[HFEM21] Lê-Nguyên Hoang, Louis Faucon, and El-Mahdi El-Mhamdi. Recommendation algorithms, a neglected opportunity for public health. *Revue Médecine et Philosophie*, 4(2), 2021.

[HFJ+21a] Lê-Nguyên Hoang, Louis Faucon, Aidan Jungo, Sergei Volodin, Dalia Papuc, Orfeas Liossatos, Ben Crulis, Mariame Tighanimine, Isabela Constantin, Anastasiia Kucherenko, et al. Tournesol: A quest for a large, secure and trustworthy database of reliable human judgments. *arXiv preprint arXiv:2107.07334*, 2021.

[HFJ+21b] Lê-Nguyên Hoang, Louis Faucon, Aidan Jungo, Sergei Volodin, Dalia Papuc, Orfeas Liossatos, Ben Crulis, Mariame Tighanimine, Isabela Constantin, El-Mahdi El-Mhamdi, Anastasiia Kucherenko, Alexandre Maurer, Mithuna Yoganathan, Felix Grimberg, Vlad Nitu, Chris Vossen, and Sébastien Rouault. Tournesol: A quest for a large, secure and trustworthy database of reliable human judgments. *ArXiV*, 2021.

[HH21a] Keach Hagey and Jeff Horwitz. Facebook says its rules apply to all. company documents reveal a secret elite that's exempt. *The Wall Street Journal*, 2021.

[HH21b] Keach Hagey and Jeff Horwitz. Facebook tried to make its platform a healthier place. it got angrier instead. *The Wall Street Journal*, 2021.

[HHHR20] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[HKJ21] Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via resampling, 2021.

[Hoa20] Lê Nguyên Hoang. Science communication desperately needs more aligned recommendation algorithms. *Frontiers in Communication*, 5:115, 2020.

[Hol21] David Holzmüller. On the universality of the double descent peak in ridgeless regression. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[HRM+18] Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *CoRR*, abs/1811.03604, 2018.

[Hut18] Matthew Hutson. Has artificial intelligence become alchemy? *Science*, 2018.

[HY21] Reinhard Heckel and Fatih Furkan Yilmaz. Early stopping in deep networks: Double descent and how to eliminate it. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[IRW+21] Huseyin A. Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. Privacy analysis in language models via training data leakage report. *CoRR*, abs/2101.05405, 2021.

[JBK+21] Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. On transferability of bias mitigation effects in language model fine-tuning. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy

Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3770–3783. Association for Computational Linguistics, 2021.

[JHG18] Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8580–8589, 2018.

[JLZ22] Haotian Ju, Dongyue Li, and Hongyang R. Zhang. Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10431–10461. PMLR, 2022.

[Joh20] Melvin Johnson. A scalable approach to reducing gender bias in google translate, 2020.

[JSS+20] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4631–4640. PMLR, 2020.

[Jul23] Camille Julienne. Uae's edge group and g42 get into natural language processing. *Intelligence Online*, 2023.

[KHJ21] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5311–5319. PMLR, 2021.

[Kit19] Matthew Kitchen. Alexa gone bad: When a.i. assistants turn on us. *The Wall Street Journal*, 2019.

[KKM+20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 2020.

[KM11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In Timos K. Sellis, Renée J. Miller, Anastasios Kementsietsidis, and Yannis Velegrakis, editors, *Proceedings of the ACM SIGMOD International Conference on Management*

of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011, pages 193–204. ACM, 2011.

[KN17] Assimakis Kattis and Aleksandar Nikolov. Lower bounds for differential privacy from gaussian width. In Boris Aronov and Matthew J. Katz, editors, *33rd International Symposium on Computational Geometry, SoCG 2017, July 4-7, 2017, Brisbane, Australia*, volume 77 of *LIPIcs*, pages 45:1–45:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[Kol63] Andrei N Kolmogorov. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 369–376, 1963.

[Kom19] Kim Komando. You're not paranoid. your phone really is listening in. *Fox News*, 2019.

[KOV15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1376–1385. JMLR.org, 2015.

[KPR17] Gary King, Jennifer Pan, and Margaret E Roberts. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review*, 111(3):484–501, 2017.

[Kri17] Sheldon Krimsky. Sugar industry science and heart disease. *Accountability in Research*, 24(2):124–125, 2017.

[KS93] Robert A Kagan and Jerome H Skolnick. Banning smoking: compliance without enforcement. *Smoking Policy: Law, politics, and culture*, 69:78–80, 1993.

[KSG03] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in P2P networks. In Gusztáv Hencsey, Bebo White, Yih-Farn Robin Chen, László Kovács, and Steve Lawrence, editors, *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, 2003*, pages 640–651. ACM, 2003.

[KSG16] Cristin E Kearns, Laura A Schmidt, and Stanton A Glantz. Sugar industry and coronary heart disease research: a historical analysis of internal industry documents. *JAMA internal medicine*, 176(11):1680–1685, 2016.

[KSU20] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2204–2235. PMLR, 09–12 Jul 2020.

[Kuc18] James Kuczmarski. Reducing gender bias in google translate, 2018.

[LC20] Heejun Lee and Chang-Hoan Cho. Digital advertising: present and future prospects. *International Journal of Advertising*, 39(3):332–341, 2020.

[Lee19] Doyoung Lee. Big data quality assurance through data traceability: A case study of the national standard reference data program of korea. *IEEE Access*, 7:36294–36299, 2019.

[Lew20] Lori Lewis. Infographic: What happens in an internet minute 2020. *AllAccess.com*, 2020.

[LGF+23] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023.

[LGV21] Shuo Liu, Nirupam Gupta, and Nitin H. Vaidya. Approximate byzantine fault-tolerance in distributed optimization. *CoRR*, abs/2101.09337, 2021.

[LJ17] Pei Li and Adam Jourdan. Chinese chatbots apparently re-educated after political faux pas. *Reuters*, 2017.

[LKK+19] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. Webuildai: Participatory framework for algorithmic governance. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):181:1–181:35, 2019.

[LL19] Pascal Lafourcade and Marius Lombard-Platet. Get-your-id: Decentralized proof of identity. In Abdelmalek Benzekri, Michel Barbeau, Guang Gong, Romain Laborde, and Joaquín García-Alfaro, editors, *Foundations and Practice of Security - 12th International Symposium, FPS 2019, Toulouse, France, November 5-7, 2019, Revised Selected Papers*, volume 12056 of *Lecture Notes in Computer Science*, pages 327–336. Springer, 2019.

[LLS21] Fanghui Liu, Zhenyu Liao, and Johan A. K. Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 649–657. PMLR, 2021.

[LM21] Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393 – 410, 2021.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, Los Alamitos, CA, USA, oct 2016. IEEE Computer Society.

[LTLH21] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *CoRR*, abs/2110.05679, 2021.

[LTLH22] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

[LW06] M Jane Lewis and Olivia Wackowski. Dealing with an innovative industry: a look at flavored cigarettes promoted by mainstream brands. *American Journal of Public Health*, 96(2):244–251, 2006.

[LYZ+21] Xiangru Lian, Binhang Yuan, Xuefeng Zhu, Yulong Wang, Yongjun He, Honghuan Wu, Lei Sun, Haodong Lyu, Chengjun Liu, Xing Dong, Yiqiao Liao, Mingnan Luo, Congfei Zhang, Jingru Xie, Haonan Li, Lei Chen, Renjie Huang, Jianying Lin, Chengchun Shu, Xuezhong Qiu, Zhishan Liu, Dongying Kong, Lei Yuan, Hai Yu, Sen Yang, Ce Zhang, and Ji Liu. Persia: An open, hybrid system scaling deep learning-based recommenders up to 100 trillion parameters. *CoRR*, abs/2111.05897, 2021.

[LZZ+17] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5330–5340, 2017.

[MACN+20] Bill Marczak, Siena Anstis, Masashi Crete-Nishihata, John Scott-Railton, and Ron Deibert. Stopping the press: New york times journalist targeted by saudi-linked pegasus spyware operator. *Citizen Lab, University of Toronto*, 2020.

[Mad10] Alexis C. Madrigal. Google: There are exactly 129,864,880 books in the world. *The Atlantic*, 2010.

[MFG+21] El Mahdi El Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[MGL+05] David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*, 2005.

[MGR21] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. Distributed momentum for byzantine-resilient stochastic gradient descent. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

[MMS19] Julia Misersky, Asifa Majid, and Tineke M Snijders. Grammatical gender in german influences how role-nouns are interpreted: Evidence from erps. *Discourse Processes*, 56(8):643–654, 2019.

[MMWMC20] Nicole Martinez-Martin, Sarah Wieten, David Magnus, and Mildred K Cho. Digital contact tracing, privacy, and public health. *Hastings Center Report*, 50(3):43–46, 2020.

[MMZ+20] Deepak Maram, Harjasleen Malvai, Fan Zhang, Nerla Jean-Louis, Alexander Frolov, Tyler Kell, Tyrone Lobban, Christine Moy, Ari Juels, and Andrew Miller. Candid: Can-do decentralized identity with legacy compatibility, sybil-resistance, and accountability. *IACR Cryptol. ePrint Arch.*, 2020:934, 2020.

[MN20] Kris McGuffie and Alex Newhouse. The radicalization risks of GPT-3 and advanced neural language models. *CoRR*, abs/2009.06807, 2020.

[MPR22] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1878–1898. Association for Computational Linguistics, 2022.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[MS99] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[MSR16] Bill Marczak and John Scott-Railton. The million dollar dissident: Nso group's iphone zero-days used against a uae human rights defender. *Citizen Lab, University of Toronto*, 24, 2016.

[MSRD18] Bill Marczak, John Scott-Railton, and Ron Deibert. Nso group infrastructure linked to targeting of amnesty international and saudi dissident. *Citizen Lab, University of Toronto*, 2018.

[MSRM+18] Bill Marczak, John Scott-Railton, Sarah McKune, Bahr Abdul Razzak, and Ron Deibert. Hide and seek: Tracking nso group's pegasus spyware to operations in 45 countries. *Citizen Lab, University of Toronto*, 2018.

[MVSS20] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE J. Sel. Areas Inf. Theory*, 1(1):67–83, 2020.

[Net16] The International Fact-Checking Network. An open letter to mark zuckerberg from the world's fact-checkers. *The Poynter Institute for Media Studies*, 2016.

[Net22] The International Fact-Checking Network. An open letter to youtube's ceo from the world's fact-checkers. *The Poynter Institute for Media Studies*, 2022.

[NGA+18] Ritesh Noothigattu, Snehalkumar (Neil) S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*

(EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 1587–1594. AAAI Press, 2018.

[NHK19] Lisa-Maria Neudert, Philip Howard, and Bence Kollanyi. Sourcing and automation of political news and information during three european elections. *Social Media+ Society*, 5(3):2056305119863147, 2019.

[NKB+20] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[NP82] Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, pages 289–297, 1982.

[NPZC19] Emmanuel Nyaletey, Reza M. Parizi, Qi Zhang, and Kim-Kwang Raymond Choo. Blockipfs - blockchain-enabled interplanetary file system for forensic and trusted data traceability. In *IEEE International Conference on Blockchain, Blockchain 2019, Atlanta, GA, USA, July 14-17, 2019*, pages 18–25. IEEE, 2019.

[NVKM21] Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[OC10] Naomi Oreskes and Erik M Conway. Defeating the merchants of doubt. *Nature*, 465(7299):686–687, 2010.

[OC11] Naomi Oreskes and Erik M Conway. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA, 2011.

[Per20] Sarah Perez. Tiktok takes down some hashtags related to election misinformation, ignores others. *TechCrunch*, 2020.

[Per23] Billy Perrigo. The new ai-powered bing is threatening users. that's no laughing matter. *Time*, 2023.

[Pet19] Nathan Pettijohn. Of course your phone is listening to you. *Forbes*, 2019.

[PFA21] Dario Pasquini, Danilo Francati, and Giuseppe Ateniese. Eluding secure aggregation in federated learning via model inconsistency. *CoRR*, abs/2111.07380, 2021.

[PGS+21] Avinash Prabhu, Dipanwita Guhathakurta, Mallika Subramanian, Manvith Reddy, Shradha Sehgal, Tanvi Karandikar, Amogh Gulati, Udit Arora, Rajiv Ratn Shah, Ponnurangam Kumaraguru, et al. Capitol (pat) riots: A comparative study of twitter and parler. *arXiv preprint arXiv:2101.06914*, 2021.

[Phi18] Chris Phillips. The golden state killer investigation and the nascent field of forensic genealogy. *Forensic Science International: Genetics*, 36:186–188, 2018.

[PKP+18] Vadim Popov, Mikhail A. Kudinov, Irina Piontkovskaya, Petr Vytovtov, and Alex Nevidomsky. Distributed fine-tuning of language models on private data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[PMB+22] Ashwinee Panda, Saeed Mahloujifar, Arjun Nitin Bhagoji, Supriyo Chakraborty, and Prateek Mittal. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 7587–7624. PMLR, 2022.

[Pow98] David M. W. Powers. Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*, 1998.

[Pro13] Robert N Proctor. Why ban the sale of cigarettes? the case for abolition. *Tobacco Control*, 22(suppl 1):i27–i30, 2013.

[PSST21] Ouri Poupko, Gal Shahaf, Ehud Shapiro, and Nimrod Talmon. Building a sybil-resilient digital community utilizing trust-graph connectivity. *IEEE/ACM Trans. Netw.*, 29(5):2215–2227, 2021.

[PZJY20] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1314–1331. IEEE, 2020.

[RGM18] Natalie Ram, Christi J Guerrini, and Amy L McGuire. Genealogy databases and the future of criminal investigation. *Science*, 360(6393):1078–1079, 2018.

[Rob18] Sarah T Roberts. Digital detritus:'error'and the logic of opacity in social media content moderation. *First Monday*, 2018.

[Rog21] Anna Rogers. Changing the world by changing the data. *CoRR*, abs/2105.13947, 2021.

[Rou22] Sébastien Louis Alexandre Rouault. *Practical Byzantine-resilient Stochastic Gradient Descent*. PhD thesis, EPFL, 2022.

[RSF21] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4384–4396, 2021.

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[San19] Destin Sandlin. Manipulating the youtube algorithm - (part 1/3) smarter every day 213, 2019.

[Sar22] Jayshree Sarathy. From algorithmic to institutional logics: the politics of differential privacy. *Available at SSRN*, 2022.

[SBBC20] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press, 2020.

[SC00] Henry Saffer and Frank Chaloupka. The effect of tobacco advertising bans on tobacco consumption. *Journal of health economics*, 19(6):1117–1137, 2000.

[Sch23] Michael Schuman. Why chatbot ai is a problem for china. *The Atlantic*, 2023.

[SCL+22] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. Selective differential privacy for language modeling. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2848–2859. Association for Computational Linguistics, 2022.

[SD21] Irene Solaiman and Christy Dennison. Process for adapting language models to society (PALMS) with values-targeted datasets. *CoRR*, abs/2106.10328, 2021.

[See21] Deepa Seetharaman. Facebook limits employee access to some internal discussion groups. *The Wall Street Journal*, 2021.

[Sei21] Philip Seib. *Information at War: Journalism, Disinformation, and Modern Warfare*. John Wiley & Sons, 2021.

[She20] Xinmei Shen. Microsoft's xiaoice chatbot to become its own company in china. *Abacus*, 2020.

[SHKR21] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on federated learning. *CoRR*, abs/2108.10241, 2021.

[SMS+21] Fnu Suya, Saeed Mahloujifar, Anshuman Suri, David Evans, and Yuan Tian. Model-targeted poisoning attacks with provable convergence. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10000–10010. PMLR, 2021.

[Soa15] Nate Soares. The value learning problem. In *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence*, 2015.

[Sol60] Ray J Solomonoff. A preliminary report on a general theory of inductive inference. Citeseer, 1960.

[Sol18] Joan E. Solsman. Youtube's ai is the puppet master over most of what you watch. *CNET*, 2018.

[Spi81] Robert L Spitzer. The diagnostic status of homosexuality in dsm-iii: a reformulation of the issues. *The American Journal of Psychiatry*, 1981.

[SSP+13] Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1374–1383. The Association for Computer Linguistics, 2013.

[SU17] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), Jan. 2017.

[SUS21] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[TF19] Aleksei Triastcyn and Boi Faltings. Federated learning with bayesian differential privacy. In Chaitanya K. Baru, Jun Huan, Latifur Khan, Xiaohua Hu, Ronay Ak, Yuanyuan Tian, Roger S. Barga, Carlo Zaniolo, Kisung Lee, and Yanfang Fanny Ye, editors, *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*, pages 2587–2596. IEEE, 2019.

[Tho76] Judith Jarvis Thomson. Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217, 1976.

[Tur50] Alan Turing. Computing machinery and intelligence. *Mind*, 1950.

[US21] Stefanie Ullmann and Danielle Saunders. Online translators are sexist – here's how we gave them a little gender sensitivity training, 2021.

[Val84] Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[VF19] Biveeken Vijayakumar and Muhammad Marwan Muhammad Fuad. A new method to identify short-text authors using combinations of machine learning and natural language processing techniques. *Procedia Computer Science*, 159:428–436, 2019.

[VHW13] Dries Vervecken, Bettina Hannover, and Ilka Wolter. Changing (s) expectations: How gender fair job descriptions impact children's perceptions and interest regarding traditionally male occupations. *Journal of Vocational Behavior*, 82(3):208–220, 2013.

[VPKG21] Ivan Vulic, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavas. Lexfit: Lexical fine-tuning of pretrained language models. In Chengqing Zong, Fei Xia, Wenjie

Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5269–5283. Association for Computational Linguistics, 2021.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[Wan17] Lidong Wang. Heterogeneous data and big data analytics. *Automatic Control and Information Sciences*, 3(1):8–15, 2017.

[Wan20] Zhang Wanqing. The ai girlfriend seducing china's lonely men. *Sixth Tone*, 2020.

[Wik21] Wikipedia. Wikipedia:size comparisons. *Wikipedia*, 2021.

[Won21] Julia Carrie Wong. How facebook let fake engagement distort global politics: a whistleblower's account. *The Guardian*, 2021.

[Woo20] Samuel Woolley. *The Reality Game: A gripping investigation into deepfake videos, the next wave of fake news and what it means for democracy*. Hachette UK, 2020.

[WPN+19] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019.

[Wri17] David Wright. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, 22(2):212–241, 2017.

[WSM+19] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[WWKTT20] Jenifer Whitten-Woodring, Mona S Kleinberg, Ardeth Thawnghmung, and Myat The Thitsar. Poison if you don't know how to use it: Facebook, democracy, and human rights in myanmar. *The International Journal of Press/Politics*, 25(3):407–425, 2020.

[WWL+22] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):1–8, 2022.

[Xu18] Yizhou Xu. Programmatic dreams: Technographic inquiry into censorship of chinese chatbots. *Social Media+ Society*, 4(4):2056305118808780, 2018.

[YAE+18] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *CoRR*, abs/1812.02903, 2018.

[YNB+21] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. *CoRR*, abs/2110.06500, 2021.

[Yor21] Jillian C York. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso Books, 2021.

[ZBH+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[ZBH+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021.

[ZKV+19] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why ADAM beats SGD for attention models. *CoRR*, abs/1912.03194, 2019.

[ZSW+19] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019.

[ZWK+21] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting few-sample BERT fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[ZWLB20] Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? *CoRR*, abs/2011.04946, 2020.

[ZXLZ15] Tianqing Zhu, Ping Xiong, Gang Li, and Wanlei Zhou. Correlated differential privacy: Hiding information in non-iid data set. *IEEE Trans. Inf. Forensics Secur.*, 10(2):229–242, 2015.

[ZYST19] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1):5:1–5:38, 2019.

[ZZBZ20] Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning. *CoRR*, abs/2009.04872, 2020.