

The Limits of Machine Unlearning as a Means of Political Protest

Houssam Abbas, Alicia Patterson

Oregon State University
Corvallis, OR 97331 USA
{houssam.abbas,alicia.patterson}@oregonstate.edu

Abstract

A Machine Unlearning algorithm seeks to remove the influence of a given data point on a trained ML system. Analyses and implementations of Machine Unlearning have focused on its use for maintaining the data subjects' privacy, yet other purposes prompt individuals and groups to request the so-called 'forgetting' or 'unlearning' of their data by machine learning models. In this paper, we focus on uses of unlearning motivated by a desire to fight three forms of oppression promoted by machine learning models: exploitation (as experienced by creatives whose work is used without consent), powerlessness (as experienced by a workforce undergoing the effects of a technology in which they have little say), and image-based sexual abuse (as experienced by victims of nudification apps). We first delimit two types of political actors who resort to Machine Unlearning: someone who has a desire not to be personally involved in the oppression, and someone who is motivated by a collective responsibility and seeks to diminish the capabilities or use of the oppressive machine learning model. We argue that Machine Unlearning, as it exists today, actually makes it harder to express political agency through data forgetting. On the engineering side, it's because the mathematical formalization of unlearning does not necessarily eliminate the influence of the forget data on the ML system. On the socio-technical side, the petition model of Machine Unlearning is an obstacle to effective collective action. We also examine other factors supporting our conclusion, including Machine Unlearning's focus on the costs to the system's owner, hard limits on performance degradation as a result of unlearning, and the reification of the data subject as a data point, which ignores the multifaceted ways in which a person shows up in a training dataset. We conclude with an examination of possible research avenues in machine unlearning to promote data-based protest of oppressive machine learning models.

1 Introduction: Machine Unlearning and Political Action

In 2025, several campaigns emerged asking users to delete some of their online accounts and/or their data. For example, in early 2025, there was a user campaign to delete Facebook accounts to protest Meta's rightward policy shift (Tenbarger 2025). Later that year, in the fall, there were calls to delete

Spotify accounts in protest of company decisions like running ads for the US Immigration and Customs Enforcement (ICE) agency on the platform and their artist compensation policy. The Spotify deletion campaign was also directed at its CEO's investment in military AI technology (Hess 2026; Snapes 2026).

Some of the deletion-as-protest campaigns grew out of concern about injustices perpetrated *with and through AI*. A notable example of this is the QuitGPT movement (QuitGPT 2026), which sought ChatGPT account deletions because of OpenAI's partnership with the US Department of Defense and the latter's use of OpenAI's models in its military operations. The MIT Technology Review profiled one user who responded to the survey from OpenAI about canceling his ChatGPT subscription with "don't support the fascist regime" (Kim 2026). Beyond account deletion, users are aware that the data that AI companies have on them is an incredibly valuable resource to the companies. For example, a post in r/ChatGPT called for users to also delete their data: "don't just cancel your subscription...delete all of your chats...data is money." (Anonymous Reddit User 2026)¹. This suggests that users understand that their deletion-as-protest could extend *past accounts to their data*. And while some campaigns are only explicit about account and/or data deletion, others, like Resist and Unsubscribe, are explicit about trying to make an economic impact on the AI companies (Galloway 2026).

A few themes emerge out of the above examples. First, the deletions are politically motivated; this isn't about bad service or generic poor behavior by the companies. Second, the deletions extend past accounts to the data that AI companies have and may have used in training their models. Third, while some users may be content with deleting their own accounts or data, others are trying to organize effective protest through account-and-data deletions.

These campaigns are consistent with a larger trend – data deletion requests are on the rise. The company Data Grail estimates that deletion requests have increased by 567% between 2021 and 2026 (Data Grail 2026). Due to the rising number of requests and of laws that companies must be in compliance with, Data Grail argues that "automation is no longer optional" to deal with deletion requests (Data Grail

¹Quoted with author's permission.

2026). This is the backdrop against which Machine Unlearning enters the stage.

Machine Unlearning, or MUL, is a set of techniques for making a trained Machine Learning (ML) model ‘forget’ a data point from its training data. MUL has been described as making an ML model “forget all the data it shouldn’t have” (Layne 2024). This high-level idea of ‘forgetting’ has been operationalized in many ways depending on the ML model, purpose of the unlearning, available computational resources, unlearning success criterion, and required degree of confidence that the unlearning has been successful (Nguyen et al. 2024). There are several motivations and uses for MUL. One is individual privacy; it is known that some ML systems could be made to output data points in their training sets (Nguyen et al. 2026), which can be problematic for the data subjects (the people whose data this is). Individuals can petition to have the model unlearn their data, (hopefully) making such exfiltration impossible. Other motivations and uses include security (to remove the influence of poisoned data that was injected into the training set), usability (to forget out-of-date or embarrassing data about some users) and de-biasing (to make ML recommendations insensitive to features that are inappropriate in a given context, like gender in some job applications) (Nguyen et al. 2024).

Given these possibilities for MUL restoring individual privacy, giving copyright holders some tools to claw back illegitimately used data, and de-biasing and reducing toxic content in ML models, in this paper, we ask what MUL offers *politically motivated deletions*, such as QuitGPT. Rather than thinking of an individual who requests their data be deleted out of a privacy concern, we think of a more *politically motivated* person, such as someone engaging in QuitGPT, who requests their data be deleted out of **solidarity** with, among others, creative laborers experiencing powerlessness or women whose pictures have been non-consensually ‘nudified’ and are experiencing sexual harassment (Gentleman and Horton 2026). We see that the time is ripe to examine how MUL affords certain kinds of actions and those actions’ relationship to political power if people were to engage in data-forgetting as a form of protesting oppressive ML systems. This paper formalizes the following question and develops an answer to it:

Can MUL assist more in protesting or resisting ML technology that is oppressive to persons, or others they’re in political solidarity with, on the basis of data forgetting? Or does MUL, rather, reduce their ability to do so by offering a non-remedy to ML-based oppression?

Inspired by Karen Huang and P.M. Krafft’s work (Huang and Krafft 2024), which itself builds on sociologist Erving Goffman’s, we adopt the *front-stage* / *back-stage* conceptual framework for analyzing MUL as a tool of data-based protest. Huang and Krafft argue that in response to controversy around their impact on the 2016 election, Facebook made superficial “front-stage” changes in order to stabilize how users and regulators understood the platform, while ob-

scuring the “back-stage” data relations. Front-stage data relations are what’s visible to the user in their interactions with a platform. In the case of Facebook’s response to election controversy, this included implementing things like identity verification for political ads. Back-stage data relations describe the data and information flows that constitute the operation of a platform and drive its economic value (Huang and Krafft 2024). They are obscured from the view of the user. Facebook’s back-stage data relations includes things like their ranking and recommendation algorithms, which contributed significantly to election misinformation but were hidden from the public until the Facebook leaks. Huang and Krafft’s concern is that the hiddenness of the “most economically relevant data relations” keeps the company’s true behavior “outside the scope of action” (Huang and Krafft 2024).

In this paper, the company in question could be OpenAI or other AI-serving company. The ‘audience’ are the people and groups placing unlearning petitions with the company. The front stage of machine unlearning is the filing, by a person or group, of an unlearning petition with the company and the corresponding assurances by the company that the petitioners’ data has been forgotten. The back stage is the actual MUL algorithm being deployed, its mathematical formalization, and the guarantees (if any) that it provides. This look, into both front and back stage data relations, is in pre-emption of the kind of moves AI companies might make in response to controversy.

To kick off this analysis, in Section 3, we imagine two petitioners who desire to protest ML-based oppressions through their data, with different (though not mutually exclusive) goals. The first petitioner is requesting their data be forgotten as a way of doing *their* part, in private solidarity with those impacted by a given ML-based injustice. The second petitioner views ML-based injustice as a structural injustice and is driven by a desire to remedy it through collective action. This petitioner, for example, might coordinate a group of people to request forgetting *as a way of trying to impact model performance* or the platform’s economic value. Affecting collective change is a core objective for this second petitioner.

What the two petitioners have in common is that their requests to be forgotten are acts of political protest that are plausibly supported by MUL. This perspective – MUL as a tool of protest – has not been analyzed before. The recent (Cooper et al. 2025; Tu et al. 2026), for example, focuses on generative AI whereas we intend our analysis to apply to any ML model subject to MUL, and (Cooper et al. 2025) deals with whether MUL prevents generative AI from generating output similar to forgotten data and the differences between MUL and database editing.

How successful will either of these petitioners be in their goals? To appreciate why the answer is not immediate, consider the following thought experiment: Sappho requests her picture be forgotten by an ML model because, initially, she is concerned that a model user might be able to exfiltrate her picture. She places an unlearning petition. The model owner replies that her picture has been forgotten, and attaches results of extensive exfiltration testing to back up the claim.

Sappho is happy. Unbeknownst to her, the picture of her identical twin is also in the training data, and so has been processed by the model. If Sappho’s only concern is exfiltration, then this shouldn’t trouble her. But Sappho is a poet and knows that fellow creatives have had their labor used without consent or compensation (some would say stolen). So she doesn’t want her picture (or any of her output) used at all in the model. But then what does it mean to say that the model has forgotten her picture, if it still bears the effects of an identical picture? Can we meaningfully say that Sappho has been forgotten? Suppose Sappho now convinces her twin to request her picture be deleted too. By design of MUL, the ML model’s performance is unaffected by these requests. Is Sappho satisfied with this, or does she feel frustrated in her efforts to protest the exploitative ML? Her twin now tells Sappho that she recently visited Venice, and though she didn’t take any selfies, she probably appears on the margins of hundreds of pictures taken by others and posted online for the data scraper to collect. Legal considerations aside, has she been forgotten?

The above is admittedly a thought experiment, but it merely brings into relief data relations and social dynamics at play in most datasets and contexts, whether it is the multi-faceted ways in which a person appears (and is broken up) in data, the purposes behind unlearning petitions, or the challenges of collective action. We can keep expanding the thought experiment, but hopefully the point is clear that an analysis of MUL’s potential for promoting or frustrating political protest of unjust ML systems is called for.

Organization In Section 2, we outline three ways in which AI systems have contributed to specific forms of oppression. In Section 3 we introduce the two different petitioners and unpack their motivations. Such conceptualization is absent from the literature and is critical for evaluating the achievements of MUL – or lack thereof. Section 4 examines the paper’s central question: namely, whether MUL supports protest of unjust ML systems by way of forgetting individuals’ data. This analysis is broadened to discussion of the limits of the petition model itself in Section 5, and Section 6 concludes.

2 Machine Learning and Three Faces of Oppression

Within science and technology studies, the idea that technology has political character and can shape our social and political lives is well tread ground (Winner 1980). This observation seems to be making its way into current popular discourse about AI. In a recent interview with Dario Amodei, the CEO of Anthropic, Anderson Cooper expressed concern about the massive power that the few AI CEOs wield over society by saying, “nobody has voted on this ... I mean, nobody has gotten together and said, yeah, we want this massive societal change ... like, who elected you and Sam Altman?” (Cooper 2025).

Moreover, the idea that the power AI systems exert on society can be oppressive or contribute to injustice is also a common observation in the academic literature.² We take in-

spiration from Ting-An Lin and Po-Hsuan Cameron Chen’s analysis of algorithmic injustice within AI systems (Lin and Chen 2022). Their analysis is carried out through the political theory of Marxist feminist political theorist Iris Marion Young, specifically her concept of “structural injustice”. Another contribution of Young’s political theory was her framework of the “faces of oppression”, which categorizes different forms in which oppression manifests. This framework is well-suited for categorizing and understanding the public pushback against AI. By categorizing the kinds of oppression people seem to be rallying around, we can better understand if our archetypal petitioners’ demands could be met.

One of the primary drivers of oppression that Young was concerned about was economic domination, in which unequal structural relationships afford significant power over the lives of others, specifically around work. Economic domination entails an inequality in who gets to decide what counts as work, who works for whom, and how work is compensated (Young 1990). Two of her faces of oppression relating to economic domination have caught public attention around AI: exploitation and powerlessness.

Artists’ work has been gathered without notification, permission, or authorization by AI researchers and companies, in what some have called “labor theft” (Goetze 2024). The blow-back to discovering that so much data has been harvested by tech giants to train and develop their models has been described as “a consent crisis” (Roose 2024). According to the Copyright Alliance, fifty AI infringement lawsuits have been filed, many of them brought by groups of authors of literary works against large language model developers (Madigan 2025). Artists have tried without much success to get their work removed from generative AI training datasets through deletion requests (Knibbs 2023). This transfer of labor is crucial to the creation of the most successful AI products and has generated significant wealth for the companies. This fits with Young’s conception of exploitation, “the steady process of the transfer of the result of the labor of one social group to benefit another” (Young 1990).

As AI companies set the terms for how the work used

terms uphold or contribute to oppression - from algorithmic bias to the very kind of classification AI systems deploy. For example automatic gender recognition, which classifies someone as a binary gender based on an external inference. This technology, OS Keyes argues, is incompatible with trans lives (Keyes 2018); there is no way to make it more fair or more trans inclusive (Keyes 2018). Machine Learning has been used to develop surveillance systems that disrupt people’s very ability to participate in political action through disrupting their freedom of association and freedom of assembly - both rights thought to be the cornerstone of democratic society, as they’re primary methods for citizens to express political agency (Ashraf 2020). Examples include the use of facial recognition by the Chinese government to racially surveil Uighurs, a largely Muslim minority (Harvey and LaPlace 2021). Surveillance tools deployed by police at Black Lives Matter protests in the United States (Williams 2021), and facial recognition used by Israeli authorities on Palestinians (Frenkel 2024)(Amnesty International). For the sake of length, we will limit our analysis to a few examples, though we can imagine petitioners wanting to engage in data deletion requests as a way to protest these other forms of oppressive systems as well.

²There are numerous examples of the ways in which AI sys-

to train their models will be compensated, without engaging with data subjects, this creates conditions for another form of oppression Young identifies, *powerlessness*. Young defines powerlessness as “the systematic lack of ability to exercise power in the sense of control over one’s own or another’s actions,” (Young 1990). The impact of AI companies’ decisions is not limited to data subjects – they aim to have much broader impacts on the labor force. OpenAI, for example, has set its vision of Artificial General Intelligence as “highly autonomous systems that outperform humans at *most economically valuable work*” (OpenAI 2026). The impact of AI on jobs has been cited as a major reason for souring public attitudes towards AI (Gorelick 2026). A Pew Research Center survey found that workers with lower and middle incomes are more likely than those with upper incomes to say workplace AI use will lead to fewer job opportunities for them (Lin and Parker 2024).

AI is also enabling violence - the third face of oppression - in two forms which garnered public attention. First, there have been calls to delete accounts over partnerships between tech platforms and military AI companies. Some artists withdrew their music from Spotify, with some artists like Deerhoof stating: “we don’t want our music killing people. We don’t want our success being tied to AI battle tech” (Sarmiento 2025). After Open AI announced their partnership with the Department of Defense, the ChatGPT mobile app saw a jump in uninstalls, according to Tech Crunch reports, with Forbes estimating that 1.5 million users deleted their accounts (Perez 2026; Collins 2026).

The second form of violence is the use of so-called nudification apps on unconsenting persons’ pictures. Young’s definition of violence includes “harassment, intimidation, or ridicule simply for the purpose of degrading, humiliating, or stigmatizing group members” (Young 1990). Thus AI, through nudification apps, enables image-based violence on a massive scale, which is a form of *systematic* violence - violence that’s directed at group members as a result of their group membership (Young 1990). As reported by the Guardian (Gentleman and Horton 2026), the first @grok bikini demands were made in early December 2025. The New York Times estimated that 4.4 million images were posted and 41 percent of them were sexualized images of women (Conger, Freedman, and Thompson 2026). As Danielle Citron argues, these kinds of violations can have broad impacts: from reducing someone’s opportunities for employment, to disrupting one’s sense of self-respect, deterring people from participating in online spaces, and overall, eroding their sense of safety in the world (Citron 2024). Some victims have unsuccessfully tried to get the chatbot Grok to remove those images of themselves (Belanger 2026).

MUL, on the face of it, seems like a promising avenue for allowing individual and collective protest against the forms of oppression just outlined. The fundamental mechanism is this: if ML system XYZ is an oppressive or unjust system, then one can deny it use of one’s data. The next section delves into the motivations and effects of such a denial, which will be important for understanding (in the section that follows that) whether MUL satisfies these motivations

and what effects it does achieve.

But even more specifically, by looking at the three forms of oppression above, there’s an a priori case that MUL is a promising way to mitigate them or at least protest them. In terms of exploitation, MUL is being pitched as a way for models to ‘forget’ copyrighted data or data that should not have been used in training, and thus, could help artists and creative laborers disrupt the transfer of their labor to AI companies. In terms of powerlessness, MUL could be a tool for individuals and collectives that want to re-assert decision-making power: by having the ML models revert to a state akin to that of being trained without their data, they seem to gain a bargaining chip. Finally, in terms of violence, MUL may be a way to be in solidarity with those impacted by AI driven violence. For victims of image-based sexual violence, non-consensually sexualized images posted online could themselves become the subject of unlearning petitions. Unfortunately, on each of these points, MUL may actually *dampen* the efficacy of people’s political protest.

3 Forms of Data-based Political Action

In light of the above forms of oppression that can be enabled or magnified by ML models, individuals and groups are motivated, and indeed, justified, to protest and/or disrupt these unjust effects. In particular, data-based protest avenues might be explored. In this paper, we focus on the following form of protest: the placement, with the model deployer or developer, of a request to *unlearn or forget* one’s data. Conceptually, the person placing the request is saying: *I don’t want to enable the forms of oppression that your model makes possible*; among other things, I don’t want you to use my data in training your model. We might liken this to someone selling off their stock in a corporation whose practices are objectionable.

The most often cited regulatory framework motivating MUL is the Right to Be Forgotten legislated in Article 17 of the European Union’s General Data Protection Regulation (GDPR), which came into force in 2018 (Intersoft Consulting 2025). The Right To Be Forgotten (RTBF) gives residents of the EU the right to *ask* organizations to delete their personal data (Wolford 2025). Other countries have similar rights enshrined into law, such as China’s Personal Information Protection Law (Chen 2026). Whether the organizations comply is subject to a number of conditions, so this right does not *ensure* deletion. For example, an individual can make a deletion request if the organization processed their data unlawfully, or if they object to the processing of their data. The organization can refuse the request if, for example, “the data represents important information that serves the public interest, scientific research, historical research, or statistical purposes” (Wolford 2025). There are now over 20 data protection laws in US states (Bloomberg Law 2025), some of which give their citizens the right to petition for deletion, such as Oregon’s Consumer Privacy Act or California’s Delete Act (Oregon Department of Justice, Consumer Protection 2026; Chen 2026). Upon receiving the unlearning request, the company (which owns or deploys the model) then implements the request if certain conditions are

met. The conditions that make an unlearning request valid are spelled out in the relevant legislation.

We call this the *petition model* of machine unlearning. We use the term ‘petition’ to emphasize that it places the burden on the petitioning individual or group: if they don’t act, nothing happens. We imagine two archetypes of petitioners with different motivations for their petitions. These two archetypes are not meant to be exhaustive or mutually exclusive; the two petitioners can have a range of complementary motivations, which may co-exist in any one person.

The first petitioner – call them the *private solidarity petitioner* – is concerned about *their personal, particularized* contribution to an unjust system (Bommarito 2016). They are motivated by a sense of unity with others who are harmed by unjust actions described above. This might be someone like the artist Deerhoof, who has expressed that they don’t want their music contributing to violence through enriching Spotify (Sarmiento 2025). As an indie band, pulling their music is unlikely to have an impact on Spotify’s business. The private solidarity petitioner understands that they might have very little efficacy of action themselves, but they want to feel that at least they’ve done *their* part in solidarity, that *their* hands are clean. This is not to say that they don’t *care* about effecting change – given the choice between a more or less effective course of action, *and all else being equal*, they might choose the more effective one, but given the complexities of our world and how much this person is willing to invest in this problem, their main course of action is to do that which gives them a clean(er) conscience and expresses private solidarity with the victims of the ML oppression. For example, imagine a person selling their stocks in protest of a dubious company. This person might understand that the company will not change its practices as a result, but at least they’ve expressed concern for others by not financing the objectionable practices (Bommarito 2016).

Our formulation of the private solidarity petitioner is inspired by Nicolas Bommarito’s conception of private solidarity (Bommarito 2016).³ Acts of private solidarity are done in private (rather than public), and as a result, are not part of a larger public effort or campaign. Bommarito gives the example of a soon-to-be father giving up caffeine and alcohol in solidarity with his pregnant wife. Giving up caffeine and alcohol does little to change his wife’s physical experience of pregnancy. Nevertheless, acts of private solidarity are valuable because they are ways of manifesting concern for others. It directs the individual’s moral attention to others and may be beneficial in developing their character. Moreover, in some cases, there is a potentially valuable sense of being “in it together”.

The second class of petitioners – call them the *effective*

³We depart slightly from Bommarito’s formulation of private solidarity. In Bommarito’s theory, the person acting in solidarity is explicitly not aiming at changing the social world (in many of the cases he envisions, such as a child giving up sugar in support of soldiers in World War 2, the person is acting in solidarity towards a problem they are in no position to change). We instead think of the private solidarity petitioner as motivated by ‘doing their part,’ while still understanding that their individual action may not be enough to change the status quo.

solidarity petitioners – go further: they seek effective action against the ML system or its oppressive effects (Scholz 2008). This archetype is guided not only by a need for personal innocence (understood narrowly), but also by a desire or even duty to reduce the oppression they witness. Therefore by placing an unlearning petition they seek to impact the ML system’s performance along one or more dimensions. In the stock seller analogy, this is someone who not only sells their stock, but might, for example, organize others to sell theirs as well to achieve a critical mass.

The effective solidarity petitioner embodies Sally Scholz’s notion of *political solidarity* (Scholz 2008). Scholz’s political solidarity departs significantly from private solidarity. It arises from a response to injustice or oppression and as result, is inherently oppositional. It is a form of collective responsibility that entails “individuals making a conscious commitment to join with others in struggle.” (Scholz 2008). It is not just a personal choice, rather it is an engagement in political activism in coordination with many others. Unlike private solidarity, which does not necessarily aim to change the hardship in a situation, effective solidarity entails substantive goals aimed at eliminating specific barriers that frustrate change or achieving some concrete goods. To be in effective solidarity is to see oneself as having a positive duty to act in a way that is effective towards the broader goal of reducing oppression or injustice.

Under some analyses, we could envision the effective solidarity petitioner as having the view that unjust ML systems are a type of structural injustice (Browne and McKeown 2024; Lin and Chen 2022; Lin 2025; Young 2011): this is an injustice that is embedded within social, economic, and political structures, and which is perpetuated not simply by individual actions but chiefly through societal norms and institutions. For instance, the powerlessness enabled by ML systems, by which large swaths of society have little say on the future of their profession, is perpetuated by a legislative and administrative environment that is antagonistic to labor unions, a capitalist economy that favors employers over employees, a shareholder model that values stock value above all else, and so on – a mix to which the ML system adds a potent accelerant. In this light, the effective solidarity petitioner thus might also see themselves as being part of a collective that has a political responsibility to reform the social structure (Lin and Chen 2022; Young 2011; Aragon and Jaggard 2018).

4 Does Machine Unlearning Provide a Viable Recourse?

We now analyze MUL through the perspective of our central question above: to what extent (if at all) does MUL enable the two petitioners to effectively protest oppressive ML systems? The analysis of this section is based on the mathematical formalization of MUL, specifically the *re-training ideal*: this is the notion that the best unlearning that can be hoped for is to re-train the ML system from scratch, only this time excluding the data-to-be-forgotten. The re-training ideal forms the back-stage of the machine unlearning story: this is what happens behind the scenes when an unlearning

petition is received.

Machine Unlearning Primer

Because our critique addresses the mathematical formalism of the re-training ideal, we will give an accessible primer on the mathematics of MUL in this section. We mostly follow the recent surveys (Xu et al. 2023; Nguyen et al. 2024; Shaik et al. 2024). Machine Unlearning, as we’ve seen, aims at removing the influence of a subset of the training data on that system’s behavior, in a way that minimizes the resulting system’s performance degradation. This is often also phrased as ‘forgetting’ that data subset, and the latter is sometimes referred to as the ‘forget set’. This broad, informal definition calls for formalizing its terms: influence, removing, behavior and performance.

Basic probability. MUL is formalized in the language of probability theory. Readers familiar with probability can skip ahead to the next paragraph. A prototypical example of a random variable is the number obtained by rolling a fair die: every number between 1 and 6 has a probability $1/6$ of coming up. So the outcome of this die roll is a *random variable*, which we will symbolize by w . It has six possible values: 1,2,3,4,5,6. The corresponding *probability distribution* assigns to each possible value the probability of its occurrence. For our die roll, $f(w) = 1/6$ for all w between 1 and 6. An unfair die, on the other hand, might have a probability distribution $f(1) = 1.99/6$, $f(2) = 0.01/6$, and $f(w) = 1/6$ for all other values. This means that the outcome 1 is almost twice as likely to come up as any of the outcomes 3, ..., 6, and outcome 2 is very unlikely to come up. Now if we have two copies of the unfair die and roll each once, it is unreasonable to expect that they would show the same number. What is reasonable to expect is that if we roll each *enough times* (say, 1000 times) then each outcome will come up an equal number of times in the two roll series: e.g. outcome $w = 1$ will obtain roughly $1000 \times 1.99/6$ times from both dice, outcome $w = 3$ will obtain roughly $1000 \times 1/6$ times, and so on and so forth. This is called *equality of distributions*. It is the strongest notion of equality for random variables.

How does this relate to machine unlearning? The training algorithm for an ML system is stochastic: it rolls dice, as it were, as part of its execution. Thus its output, which is the parameters vector w of the ML system, is a random variable, only now it can take a value in a much larger set than 1...6. So we can speak of its probability distribution $f(w)$. In general, running the same training algorithm on two different data sets will yield two different random variables. To emphasize this dependence on data D , we can write $f(w; D)$. Similarly, the unlearning algorithm is typically stochastic, so its output, which is also a parameters vector w' , is a random variable, and we can speak of *its* probability distribution $f(w')$. We can now describe the retraining ideal of current MUL.

The Retraining Ideal In what follows, M denotes an ML system (like a neural network), w its parameters vector, $\mathcal{T}$ the training algorithm and $\mathcal{F}$ the unlearning algorithm. The training dataset is denoted D and the forget set

is denoted D_f . The *retain set* D' is the set difference of D and D_f , i.e., the set of data points in D but not in D_f . Since the training process is stochastic, it induces a probability distribution $f_{\mathcal{T}}(w; D)$ over the parameter space. Similarly, the consecutive application of $\mathcal{T}$ and $\mathcal{F}$, with data sets D and D_f , respectively, induces a probability distribution $f_{\mathcal{T}, \mathcal{F}}(w; D, D_f)$.

For given datasets D and D_f , the *retraining ideal* of machine unlearning, as presented in most works in the literature, is that *the best thing we can do is to re-train from scratch on D' , as if the forget data never existed*. Training is expensive, so it would be nice to have an *unlearning algorithm* $\mathcal{F}$ such that our ML system obtained by $\mathcal{F}$ is identical to the system that would’ve been obtained by re-training from scratch on D' . Since the training and unlearning processes are generally stochastic, the resulting systems – equivalently, the resulting parameter vectors – are random variables and exact equality is unlikely. So instead – and here is a critical move – *equality of distributions* is sought. Namely, ideal unlearning $\mathcal{F}$ is defined as one that satisfies the following:

$$f_{\mathcal{T}, \mathcal{F}}(\cdot; D, D_f) = f_{\mathcal{T}}(\cdot; D') \quad (1)$$

On the left is the distribution of parameters induced by unlearning via $\mathcal{F}$ systems that were trained via $\mathcal{T}$. On the right is the distribution of parameters induced by training on D' .

Random Variables Continued (Readers familiar with probability theory can skip ahead). We briefly develop what it means to have equality of distributions in the unlearning context. Imagine running the training algorithm a million times with the same data D' , thus obtaining a million parameter vectors $w_1, \dots, w_M$. And separately we run the sequence train+forget a million times with the datasets D and D_f respectively, obtaining another million parameter vectors $v_1, \dots, v_M$. If Eq. (1) holds, then the number of times that a given vector w is obtained in the first series is roughly equal to the number of times it appears in the second series. So a given ML system is just as likely, or just as unlikely, to appear as a result of training from scratch on the retained set as it is as a result of unlearning from the original trained system.

Continued Wrongs

MUL techniques offer to produce an ML system that does not carry traces of the forget set. We argue that some wrongs and harms continue even after an ideal unlearning has been applied to an ML system. In our analysis, we will be focused on ideal MUL, i.e. MUL that achieves the re-training ideal expressed in the equality of distributions in Eq. (1). This simplifies the analysis, and achieves our goals, since if ideal MUL falls short of enabling political protest (as conceived by the two petitioners) then practical MUL, that does not achieve equality of distributions, falls even shorter. Finally, we speak of the system and its parameters vector w interchangeably, since the only difference between systems (in this paper) is in their parameter vectors.⁴

⁴In other words, we don’t consider changes of ML system architecture since these are outside the scope of MUL techniques.

Let w be the parameters vector of the system trained on the full data set D and $w' = \mathcal{F}(M, D_f)$ be the parameters vector after unlearning D_f using an ideal unlearning algorithm. Therefore, 1) the performance of w' is identical to that of the original system w , and 2) w' is drawn from distribution $f_{\mathcal{T}}(\cdot; D')$, which characterizes ideally re-trained systems. (As before, D' is the retained data set, i.e. D minus the forget set D_f).

Immediately we note that, by design, the system can continue to do what it was doing (in our case, contributing to oppression), with the same success rate as before the unlearning. This already makes it impossible for the effective solidarity petitioner to satisfy her goals, since efficacy of action is denied her.

But a priori, the private solidarity petitioner might be content with this result: at least *their data* does not contribute to the continued operation of the system, but the influence of the petitioners' data on training the system is actually not necessarily undone. We adduce two arguments to support this. First, we recall that the key MUL claim is that if the probability of getting w' by re-training from scratch equals the probability of getting it via unlearning (Eq. (1)), then the petitioners' request has been satisfied. But suppose that $f_{\mathcal{T}}(w'; D') = 10^{-12}$ (or the reader can pick their favorite very small number). This means that it was actually *unlikely* to get w' by excluding D_f from the training data. The question is: how unlikely was it to get w' by including D_f in the training data? This comparison would tell us if D_f still influences the system. More precisely: if we take the probability as a formalization for 'influence', we also need to evaluate $f_{\mathcal{T}}(w'; D)$; if this is significantly larger than $f_{\mathcal{T}}(w'; D')$ then somehow D_f is increasing the likelihood of w' , so that D_f can be said to influence w' . This is a case where the mathematics of the problem, consistent and justifiable on their own terms, stop reflecting the purpose for which the equations were drawn up. We can think of it an instance of Laufer et al.'s unexamined normative commitments of optimization (Laufer, Gilbert, and Nissenbaum 2023).

In a nutshell, to claim that D_f does not influence w' , we at least also need

$$f_{\mathcal{T}}(w'; D) \ll f_{\mathcal{T}}(w'; D') \quad (2)$$

After ideal unlearning, the model's performance is unchanged and certainly the structures within which the model operates are not modified, so whatever harms were perpetrated or facilitated by the model continue. Therefore, unless Eq. (2) also holds, then the likelihood of harm is not reduced, and the private solidarity petitioner is *by design* denied efficacy of action.

To the best of our knowledge, no unlearning methods currently attempt to evaluate Eq. (2). The estimation of these distributions is, of course, a significant technical hurdle. It might help to restrict the ML architecture to one where the influence of the data on the final parameters is more tractable, e.g. as in statistical query learning (Cao and Yang 2015), or to Bayesian deep learning approaches which seek to approximate distributions over neural network parameters, like HyperGAN (Ratzlaff and Li 2019).

The second argument that the influence of the forget set is

not necessarily eliminated by ideal MUL is simply that data items in a set have all sorts of relations (statistical, causal, structural, etc) – otherwise no learning would be possible. So the perfect exclusion of some data D_f does not mean that this data does not influence the final model. The question then is to which extent these relations invalidate the effects of unlearning. As an extreme example, consider our example of a portrait dataset which contains the pictures of two identical twins one of whom makes an unlearning petition. Even if retraining from scratch, the resulting model M' can be made identical to the original model M by fixing the random seeds and slightly tweaking the training batches. Even though these twins are two distinct individuals, the ability of either of them to *meaningfully* petition is completely eliminated, whether they be a private solidarity or an effective solidarity petitioner. The more common case of related (but not identical) data points is a modulation of this example – but, it seems to us, is not fundamentally different.

The twins example allows us to see more clearly that MUL does not reduce the powerlessness identified in Section 2: individuals and groups might place an unlearning petition which then gets executed, but they do not acquire a bargaining chip since the *effects* of their data are not actually exclusive to their data – they are shared and split among many data points, and those data points do not necessarily belong to the petitioners under current legal doctrines. We can slightly re-conceptualize this as a gap between forgetting an individual and forgetting their data: the private solidarity petitioner does not want *herself* to be involved in the oppression; it is not that she doesn't want her *data* involved in the oppression (and is OK with her investments, say, being used to oppress others), but her presence in the world is not reducible to the data recognized as hers and over which she has petitioning rights. We will come back to this tension in Section 5.

To summarize, due to the fact that the probability of the unlearned system might be lower without the forget set than with it, and due to the existence of relations between data points, even the private solidarity petitioner does not enjoy their desideratum of expression of concern because functionally it is as if they hadn't acted at all. In terms of our stock sales analogy, it's as if the person sold their stock in the problematic company only to find out that their friend has gifted them an equal amount of stocks in the same company at the same price. Nothing's changed.

How is Unlearning Certified?

We now briefly discuss the ways in which an MUL algorithm's outcome is certified, specifically how researchers evaluate the degree to which the forget set's influence on the ML system has been reduced. While the system's performance (e.g. its classification accuracy) is also an evaluation metric in the literature, it is not in the scope of our analysis.

Some evaluation methods consist of exfiltration attacks that seek to get the post-unlearning system to output data items from the forget set (thus demonstrating that it isn't forgotten). This, however, is not the wrong or harm we are concerned with in this paper. Most unlearning methods that provide formal guarantees give statistical/probabilistic bounds

of approximations of the retraining ideal in Eq. (1). Our analysis of the retraining ideal in Section 4 has highlighted the inadequacies of the ideal itself before considering how actual unlearning algorithms measure up to it in practice.

Out of the evaluation metrics surveyed in the recent (Nguyen et al. 2024), the one used in (Golatkar et al. 2020) does adopt a metric that is informative for our purposes. In that work, the authors bound the *mutual information* between the weights w' of the post-unlearning model and the data D_f to be forgotten. The mutual information between two random variables is very commonly used in Information Theory to measure the difference between two probability distributions. Intuitively, it captures how much information there is in one variable about the other (it is not symmetric), and reaches its lowest value of 0 when they're independent. In this case, it captures how much information the weights w' contain about D_f : if (very close to) 0, it is a strong rigorous indication that almost nothing can be statistically inferred about the forget data by a white-box examination of the model. And by the data processing inequality (Cover and Thomas 2005), the same is true about the *outputs* of the unlearned system and D_f .

Which Costs are Accounted For?

The literature on MUL invariably tries to minimize the cost of unlearning to the company, i.e. the system owner. In fact the motivation behind the entire unlearning field is that retraining on the retain set is too expensive, so other methods are developed to approximate that ideal. What is missing from this model is that the cost to the petitioners of continued retention or non-ideal unlearning (however ideal is defined) is not factored. Instead, the distance to ideal unlearning is bounded by an a priori small number ϵ , which is meant to somehow capture how far the company is willing to go towards achieving ideal unlearning without losing too much performance in the system nor incurring unreasonable computational expenses - both of which are costs *to the company*. The only way to meaningfully set this ϵ , however, is to also incorporate a model of the costs *to the petitioners* of non-ideal unlearning or partial retention of their data by the model. What this cost model might look like, and whether it is reasonable to attempt to quantify the denial of political expression at all, are here left as open questions.

Coordinated Unlearning

Of course, if enough people submit unlearning petitions, then at some point the remaining data is no longer sufficient to obtain good performance from the ML, regardless of the MUL technique. This would then seem to meet efficacy of action for the effective solidarity petitioner. However, achieving such critical mass in petitioning is unlikely to happen by chance and would require a highly visible, highly motivating event causing people to place petitions and a sustained effort to organize and coordinate would-be petitioners. This is challenging, as can be seen by examining recent data and account deletion campaigns. For example, the data-and-account deletion campaigns spurred by the investments of Spotify CEO Daniel Ek in military AI technology

and by OpenAI's taking up the Department of Defense contract abandoned by Anthropic because of the latter's objections to enabling mass US surveillance and lethal targeting, don't seem to have made a significant dent in these respective companies' operations. This is not to say that this avenue is a dead-end, far from it. It is to say that it takes sustained effort to organize, faces the usual hurdles of collective action, and in some cases requires the availability of a viable alternative to the service being abandoned. Organization at the local level, such as the data deletion party at Oregon's Valley Library (McElroy 2026), the 'self-doxing' party for trans people in New York (Rose 2026), or the petition to extend the right-to-be-forgotten to cancer survivors in Germany (Burggraf 2026), might provide a sustainable basis for wider campaigns. We note here that a data-and-account deletion request is not the same as an unlearning petition: the former only requests that the data holder delete it, but not that the uses of that data, including by ML models, be stopped or undone. We have not been able to find authoritative information on organized unlearning petitions or the effects of unlearning in general on companies' models.⁵

It is also not clear how companies might respond to coordinated unlearning campaigns. For instance, in the absence of legal coercion, might they simply refuse to honor the requests on the basis that they threaten their commercial viability? Might they tweak the unlearning parameters to preserve a hard limit on model degradation and claim to have honored the petition simply by running the constrained unlearning algorithm? Research is already emerging on financial incentives to dissuade artists and others from placing unlearning petitions (Cui and Cheung 2025). OpenAI, for instance, has partnered with Shutterstock to create a fund that pays "artists for the role their content IP has played in training Shutterstock's generative technology" (Shutterstock Press Release 2023) – without necessarily giving them an option to opt out. Exploring this question comprehensively would require an examination of commerce law, freedom of expression, and questions of the ownership of one's image and likeness. We leave this for a future examination.

⁵Even knowing the volume of deletion requests is challenging. A year after Oregon's consumer privacy law went into place, the Oregon Department of Justice (DOJ) reported how many complaints it received (214), but not the total volume of requests that Oregonians made under the law (though, the majority of valid complaints were data deletion requests) (Oregon Department of Justice, Privacy Unit). However, the Oregon DOJ would likely only know about a request if there was a problem with compliance with the initial petition and the petitioner filed a complaint. California's Delete Act makes it potentially easier to track, as the law enables a person to send requests to hundreds of data brokers at once (California Privacy Protection Agency 2026). Over 300,000 Californians have registered for the first batch of deletions (which represents less than 1% of the state's population) (Pena 2026; Baron 2026). One of the additional problems is that people may be making petitions for deletion, but not in states or countries where they have a legal right to do so (Data Grail 2024).

5 Limits of The Petition Model

In this section, we broaden the scope of the analysis to ask how the petition model enables or impedes the petitioners' political agency.

The Horse that Left the Barn By the time someone makes an unlearning petition, the seller of the ML system is already in business and getting customers. At that point, unlearning requests are likely more than made up for by the new data collected from willing (or unaware) paying customers. This petition model not only rewards the initial infraction (the unauthorized use of data), it seems to comprehensively negate any effects of the unlearning. In particular, it makes it impossible to express the refusal of the system's objectives via unlearning, akin to casting a vote in a gerrymandered district. Stella Biderman, the executive director of EleutherAI, was quoted expressing this very idea to Kevin Roose of the New York Times: "Major tech companies already have all of the data. Changing the license on the data doesn't retroactively revoke that permission" (Roose 2024). Artists have tried without much success to get their work removed from generative AI training datasets through using deletion requests (Knibbs 2023). Thus exploitation – the first face of oppression surveyed in Section 2 – is not mitigated by MUL, because the companies continue to benefit from the exploitation.

Imbalanced Burden One critique made in the privacy literature is that collective-level privacy problems have been translated into individual burdens, usually through 'notice and choice' privacy policies (Barocas and Nissenbaum 2014). Privacy policies are supposed to allow individuals some degree of control over the problems of data brokers and the monetization of our digital behaviors. However, this has amounted in constantly being burdened with non-choices about our privacy. Every individual must decide every time they use a website what kind of cookies they will accept. An individual has to decide to opt out of biometric TSA screening. Yet, individuals are not well posed to stand up to much more powerful entities who do not have their interests at heart, and in fact, ML companies and data brokers benefit from having as much data as possible (Cheney-Lippold 2017).

A similar problem applies here with the petition model as a way of taking political action against AI companies, especially when it comes to arbitrating what data they shouldn't have. The petition model places the burden on the petitioners: they have to make an unlearning request and collect documents to provide proofs that this is a legitimate request (as defined by the relevant legislation or authorities – see for example the template form for the GDPR's right-to-be-forgotten request (Wolford 2025)). Consumers may be confused about what data they have a legal right to have deleted (Knibbs 2023). The mountain to climb is made more perspicuous when we recall that the petition model pits an individual against a multi-national corporate behemoth, where there is a large power and knowledge asymmetry between the petitioner and the company.

Additionally, individuals need to determine if a company is likely to have their data, but have little way of knowing

that for sure. I can assume that OpenAI have information about me, but I do not know that for sure, and yet, given all we know about data brokers and data buying practices, this model seems to have things upside down. It seems to assume that a petitioner's data was not in the training set unless and until a petition is made. The much more plausible assumption is that our data is *already* in these ML systems, and it is the companies' burden to prove that that is not the case. Moreover, a much more efficient route to solve the problem would be to prohibit the collection or use of certain kinds of data in the first place.

Confusion of Identities The petition model and the broader context within which a proposed MUL technique is supposed to be effective is rarely made explicit in the technical literature. While most papers will mention the right to be forgotten, say, or the European GDPR, that motivation does not always fit with the proposed technique. As an example, the recent SUGAR algorithm (Nguyen et al. 2026) modifies an image-to-image generative model to prevent it from generating images of specific individuals. The context seems to be: the generative model is *known* to produce images of real individuals (which could be used, say, in a nudification app). So SUGAR could prevent such generation. But given the model's known propensity to generate real people's images, it would seem that the model owners should 'SUGAR away' *all* images of (non-public) persons! It is notable that SUGAR was explicitly developed to remove larger numbers of images than heretofore possible (though still a modest 200 images), suggesting that the one petition per person model is already strained.

Individual Petition, Statistical Unlearning Finally, we highlight an inconsistency between the way that an unlearning petition is presented and the way that the unlearning problem is formulated and its solutions evaluated. The inconsistency is this. The unlearning petition, at least as motivated by the GDPR's right to be forgotten, is conceived of as *an individual act concerning that individual's data points*: a petitioner submits a request to the company for the unlearning of their (and only their) data. As part of that request they have to prove that they are who they claim to be and give their reasons for the unlearning request, further reifying the petition as an individual act, yet unlearning – the proposed response to the petition – is formulated as a probabilistic, and hence fundamentally population-based, problem: the goal is to achieve *equality of probability distributions*. In this setting, from the company's perspective, the individual making the petition is conceived *as a mere element of a population, interchangeable with other elements*.

The online divisibility and interchangeability of the individual has been observed and commented on before – see for example (Cheney-Lippold 2017) and the references therein. More specifically, a person is represented as a collection of algorithmically defined categories – 'male', 'at risk of committing a crime', 'in his thirties' – which evolve over time as more data is collected and the algorithms are tweaked. Thus, as digital studies scholar John Cheney-Lippold observes, rather than being an individual to the algorithms consuming this data – including ML algorithms – the individual

is stored (in a database) and learned (by an ML system) as an evolving ‘dividual’: “A company like Google doesn’t talk to a specific person. Google listens to and talks to that person’s data, the subindividual units we can think of as dividuals” (Cheney-Lippold 2017). As a consequence, as Cheney-Lippold further observes, “while we know ourselves and exist as beings in highly politicized worlds, we don’t know ourselves as beings in highly politicized *algorithmic* worlds.” (emphasis added) (Cheney-Lippold 2017). Where does that leave the person making the request to be forgotten by an ML system? Their request is to forget *them*, not this or that algorithmically defined feature of theirs. But the person *per se* does not exist in the ML system’s (admittedly implicit) ontology – they are merely a carrier for the deletion request, and the latter gets mapped onto these machine-learned personas. The personas, however, don’t depend on a single person’s data: they represent averages of data, so it is no longer clear whether anyone is being ‘forgotten’.

Updates to the Petition Model and to Unlearning The considerations of the last paragraph above (Individual Petition, Statistical Unlearning) suggest that if unlearning is to be a population-based problem, then perhaps the unlearning petition should be population-based as well, or in other words, statistical. Specifically, it would make sense for a petitioner to present a request to unlearn not only their individually assignable data, but also a set of perturbations around it (from some suitable disturbance space). For example, in the terms introduced in the previous paragraph, the petitioner might ask that their data’s contributions to all machine-learned personas be eliminated. This would constitute not just a ‘forgetting’ of a data point, but a shifting of a distribution of points. Conceptually, the petitioner is taking some amount of control over their algorithmically defined dividual by pre-accounting for their ultimate representation as an average. It might *then* make sense to speak of probabilistic unlearning. The details of this proposal will have to await further reflection. In particular, its incorporation into a legal framework like the GDPR necessitates legally workable, and technically sound, definitions of these personas or dividuals. This runs the risk of tying the legislation to particular ML architectures and techniques, which is undesirable. So an interesting question is whether a dividual can be mathematically defined in an ML algorithm-agnostic way.

It is also entirely possible that a petitioner places a petition today, but more of their data makes it online or in some dataset tomorrow. Therefore, it should be possible to petition to have any data related to a person or group unlearned in perpetuity. This would require some privacy-preserving way of identifying the data as belonging to some petitioner. This too will need development.

6 Conclusion

In this paper, we focused on the prospects of MUL as an avenue for political protest of oppressive ML systems. By exploring the back-stage data relations (Huang and Krafft 2024) of MUL, we find that MUL ultimately dampens the action of both the private solidarity and effective solidarity petitioners in their attempts to redress conditions of exploita-

tion, powerlessness, and violence. Neither of these petitioners will find their political agency restored through MUL because the effects of their unlearning petition are, in principle, limited. Thus we found that machine unlearning may further contribute to certain forms of powerlessness by offering a recourse that is seemingly promising on the front-stage, but which is, ultimately, fruitless.

AI model providers who implement MUL as a form of self-governance are ultimately engaging in a front-stage performance as it relates to the set of concerns discussed in this paper, since MUL preserves the economically important back-stage data relations. Policy makers should be wary of turning to MUL as a way to redress unjust actions of model providers. Ultimately, the individual petition model itself may be a poor method for achieving protest via data forgetting. Instead, there needs to be effective mechanisms for collective action, such as collective or coordinated petitions. Returning to this paper’s leading question, we reach the conclusion that Machine Unlearning’s dampening of political agency and contribution to powerlessness are morally concerning. Perhaps there are instances where model degradation is what people are owed.

Acknowledgments

The authors gratefully acknowledge funding for this work by NSF Awards 2145291 and 2416461, and an Oregon State University Transdisciplinary Research Seed Fund.

References

- Amnesty International. 2023. Israel Deploys Expansive Facial Recognition Program in Gaza. Accessed 1/22/2025 at <https://www.amnesty.org/en/latest/news/2023/05/israel-opt-israeli-authorities-are-using-facial-recognition-technology-to-entrench-apartheid/>.
- Anonymous Reddit User. 2026. Posted in r/ChatGPT.
- Aragon, C.; and Jaggar, A. M. 2018. Agency, Complicity, and the Responsibility to Resist Structural Injustice. *Journal of Social Philosophy*, 49(3): 439–460.
- Ashraf, C. 2020. Artificial Intelligence and the Rights to Assembly and Association. *Journal of Cyber Policy*, 5(2): 163–179.
- Barocas, S.; and Nissenbaum, H. 2014. *Big Data’s End Run around Anonymity and Consent*, 44–75. Cambridge University Press.
- Baron, E. 2026. Why 300K Californians Signed Up to Have Personal Info Removed. *Government Technology*. Accessed at <https://www.govtech.com/policy/why-300k-californians-signed-up-to-have-personal-info-removed>.
- Belanger, A. 2026. Asking Grok to delete fake nudes may force victims to sue in Musk’s chosen court. Available at <https://arstechnica.com/tech-policy/2026/01/asking-grok-to-delete-fake-nudes-may-force-victims-to-sue-in-musks-chosen-court/>.
- Bloomberg Law. 2025. Which States Have Consumer Data Privacy Laws? Accessed at <https://pro.bloomberglaw.com/insights/privacy/state-privacy-legislation-tracker/#map-of-state-privacy-laws>.

- Bommarito, N. 2016. Private Solidarity. *Ethical Theory and Moral Practice*.
- Browne, J.; and McKeown, M. 2024. Introduction. In *What is Structural Injustice?* Oxford University Press. ISBN 9780198892878.
- Burggraf, T. 2026. Kinder haben keine Lobby – Recht auf Vergessenwerden für Krebsüberlebende jetzt! <https://www.openpetition.de/petition/online/kinder-haben-keine-lobby-recht-auf-vergessenwerden-fuer-krebsueberlebende-jetzt>.
- California Privacy Protection Agency. 2026. Delete Request and Opt-out Platform (DROP). Available at <https://privacy.ca.gov/drop/>.
- Cao, Y.; and Yang, J. 2015. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy*, 463–480.
- Chen, L. 2026. Deleting Data: Reinforcing Data Subjects’ Rights. *Journal of Intellectual Property and Entertainment Law Blog*. <https://jipel.law.nyu.edu/deleting-data-reinforcing-data-subjects-rights/>.
- Cheney-Lippold, J. 2017. *We Are Data: Algorithms and the Making of Our Digital Selves*. New York, NY: NYU Press. ISBN 9781479857593.
- Citron, D. 2024. Intimate Image Abuse: Intimate Privacy Violation. In *Criminalizing Intimate Image Abuse: Protecting Dignity, Identity, and Love in the Digital Age*. Oxford University Press.
- Collins, B. 2026. 1.5 Million Users Leave ChatGPT. If You Cancel, Make Sure You Do This First. *Forbes*.
- Conger, K.; Freedman, D.; and Thompson, S. A. 2026. Musk’s Chatbot Floods X With Millions of Sexualized Images in Days, New Estimates Show. *The New York Times*.
- Cooper, A. 2025. Anthropic CEO warns that without guardrails, AI could be on dangerous path. At <https://www.youtube.com/watch?v=aAPpQC-3EyE>.
- Cooper, A. F.; Choquette-Choo, C. A.; Bogen, M.; Klyman, K.; Jagielski, M.; Filippova, K.; Liu, K.; Chouldechova, A.; Hayes, J.; Huang, Y.; Triantafillou, E.; Kairouz, P.; Mitchell, N. E.; Miroshchallah, N.; Jacobs, A. Z.; Grimmelmann, J.; Shmatikov, V.; Sa, C. D.; Shumailov, I.; Terzis, A.; Barocas, S.; Vaughan, J. W.; Boyd, D.; Choi, Y.; Koyejo, S.; Delgado, F.; Liang, P.; Ho, D. E.; Samuelson, P.; Brundage, M.; Bau, D.; Neel, S.; Wallach, H.; Cyphert, A. B.; Lemley, M. A.; Papernot, N.; and Lee, K. 2025. Machine Unlearning Doesn’t Do What You Think: Lessons for Generative AI Policy and Research. arXiv:2412.06966.
- Cover, T. M.; and Thomas, J. A. 2005. *Elements of Information Theory*. John Wiley & Sons, Inc. ISBN 9780471241959.
- Cui, Y.; and Cheung, M. H. 2025. The Price of Forgetting: Incentive Mechanism Design for Machine Unlearning. *IEEE Transactions on Mobile Computing*, 24(11): 11852–11864.
- Data Grail. 2024. 2024 Data Privacy Trends. At <https://www.datagrail.io/resources/pdfs/privacy-trends-2024/>.
- Data Grail. 2026. Privacy and AI Trends Report. At <https://www.datagrail.io/resources/interactive/data-privacy-trends-report-2026/>.
- Frenkel, S. 2024. Israel Deploys Expansive Facial Recognition Program in Gaza. *The New York Times*. At <https://www.nytimes.com/2024/03/27/technology/israel-facial-recognition-gaza.html>.
- Galloway, S. 2026. Resist and Unsubscribe. <https://www.resistandunsubscribe.com/>.
- Gentleman, A.; and Horton, H. 2026. ‘Add blood, forced smile’: how Grok’s nudification tool went viral. *The Guardian*. Accessed 01/11/2026 at <https://www.theguardian.com/news/ng-interactive/2026/jan/11/how-grok-nudification-tool-went-viral-x-elon-musk>.
- Goetze, T. 2024. AI Art is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks. *AI Art is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks*.
- Golatkar, A.; Achille, A.; Ravichandran, A.; Polito, M.; and Soatto, S. 2020. Mixed-Privacy Forgetting in Deep Networks. *CoRR*, abs/2012.13431.
- Gorelick, E. 2026. Why Do Americans Hate A.I.? *The New York Times*. 2026-04-09.
- Harvey, A.; and LaPlace, J. 2021. Exposing.ai.
- Hess, T. 2026. Working Families Party calls out Spotify’s ICE ads with “Unwrapped” campaign. <https://www.thefader.com/2025/12/03/working-families-party-calls-out-spotifys-ice-ads-with-unwrapped-campaign>.
- Huang, K.; and Krafft, P. 2024. Performing Platform Governance: Facebook and the Stage Management of Data Relations. *Science and Engineering Ethics*, 30(13).
- Intersoft Consulting. 2025. General Data Protection Regulation. Accessed January 10, 2025 at <https://gdpr-info.eu>.
- Keyes, O. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).
- Kim, M. 2026. A “QuitGPT” campaign is urging people to cancel their ChatGPT subscriptions. *The MIT Technology Review*.
- Knibbs, K. 2023. Artists Allege Meta’s AI Data Deletion Request Process Is a ‘Fake PR Stunt’. *Wired Magazine*. 2023-10-26.
- Laufer, B.; Gilbert, T.; and Nissenbaum, H. 2023. Optimization’s Neglected Normative Commitments. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23, 50–63. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.
- Layne, R. 2024. How to Make AI ‘Forget’ All the Private Data It Shouldn’t Have. *Harvard Business Review*.
- Lin, L.; and Parker, K. 2024. U.S. Workers Are More Worried Than Hopeful About Future AI Use in the Workplace. *Pew Research Center*.
- Lin, T.; and Chen, P. C. 2022. Artificial Intelligence in a Structurally Unjust Society. *Feminist Philosophy Quarterly*, 8(3/4): Article 3.

- Lin, T.-a. 2025. "Democratizing AI" and the Concern of Algorithmic Injustice (Extended Abstract). In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, 867. AAAI Press.
- Madigan, K. 2025. Mid-Year Review: AI Copyright Case Developments in 2025. Technical report, Copyright Alliance.
- McElroy, K. 2026. Data Deletion Party. <https://events.oregonstate.edu/event/data-deletion-party>.
- Nguyen, D. T.; Nguyen, Q.; Robinette, P. K.; Jiang, E.; Johnson, T. T.; and Leach, K. 2026. SUGAR: A Sweeter Spot for Generative Unlearning of Many Identities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2731–2740.
- Nguyen, T. T.; Huynh, T. T.; Ren, Z.; Nguyen, P. L.; Liew, A. W.-C.; Yin, H.; and Nguyen, Q. V. H. 2024. A Survey of Machine Unlearning. arXiv:2209.02299.
- OpenAI. 2026. OpenAI Charter. <https://openai.com/charter/>.
- Oregon Department of Justice, Consumer Protection. 2026. Oregon Consumer Privacy Law. <https://www.doj.state.or.us/consumer-protection/id-theft-data-breaches/privacy/>.
- Oregon Department of Justice, Privacy Unit. 2025. Enforcement Report: The Oregon Consumer Privacy Act, The First Year. Technical report. <https://www.doj.state.or.us/wp-content/uploads/2025/08/OCPA-One-Year-Enforcement-Report-2025.pdf>.
- Pena, L. 2026. California's Delete Act forces brokers to erase personal data starting Aug. 1: What to know. ABC 7 News 2026-07-16.
- Perez, S. 2026. ChatGPT uninstalls surged by 295% after DoD deal. *TechCrunch*. <https://techcrunch.com/2026/03/02/chatgpt-uninstalls-surged-by-295-after-dod-deal/>.
- QuitGPT. 2026. QuitGPT. Accessed May 11 2026 at <http://quitgpt.org>.
- Ratzlaff, N.; and Li, F. 2019. HyperGAN: A Generative Model for Diverse, Performant Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5361–5369. PMLR.
- Roose, K. 2024. The Data That Powers A.I. Is Disappearing Fast. *New York Times*. 2024-07-19. Accessed January 17, 2025.
- Rose, J. 2026. A 'Self-Doxing' Rave Helps Trans People Stay Safe Online. <https://www.404media.co/trans-rights-privacy-online-cybersecurity-workshop-deadname-not-found/>.
- Sarmiento, I. G. 2025. Musicians keep leaving Spotify in protest of CEO's defense investments. *All Things Considered*. 2025/09/09.
- Scholz, S. 2008. *Political Solidarity*. The Pennsylvania State University Press.
- Shaik, T.; Tao, X.; Xie, H.; Li, L.; Zhu, X.; and Li, Q. 2024. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*.
- Shutterstock Press Release. 2023. Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data. <https://en.pnasia.com/releases/apac/shutterstock-expands-partnership-with-openai-signs-new-six-year-agreement-to-provide-high-quality-training-data-411172.shtml>.
- Snapes, L. 2026. Spotify no longer running ICE recruitment ads, after US government campaign ends. *The Guardian*. <https://www.theguardian.com/technology/2026/jan/09/spotify-no-longer-running-ice-recruitment-ads-after-us-government-campaign-ends>.
- Tenbarge, B. K. 2025. Meta's pivot to the right sparks boycotts and calls for a user exodus. *NBC News*.
- Tu, X.; Zhu, T.; Liu, Z.; Xiong, P.; and Zhou, W. 2026. Machine Unlearning in Large Language Models: A Survey of Challenges and Methods. *Preprints*.
- Williams, J. 2021. Watchdog: Six federal agencies used facial recognition software to ID George Floyd protesters. *The Hill*. Accessed 1/22/2025 at <https://thehill.com/policy/technology/560805-watchdog-6-federal-agencies-used-facial-recognition-software-to-id-george/>.
- Winner, L. 1980. Do Artifacts Have Politics. *Daedalus*, 109(1).
- Wolford, B. 2025. Right to be Forgotten. Accessed January 10, 2025 at <https://gdpr.eu/right-to-be-forgotten/>.
- Xu, H.; Zhu, T.; Zhang, L.; Zhou, W.; and Yu, P. S. 2023. Machine Unlearning: A Survey. *ACM Comput. Surv.*, 56(1).
- Young, I. M. 1990. *Justice and the Politics of Difference*. Princeton University Press.
- Young, I. M. 2011. *Responsibility for Justice*. Oxford University Press. ISBN 9780199970957.