

Eternal Sunshine of the Mechanical Mind: The Irreconcilability of Machine Learning and the Right to be Forgotten

Meem Arafat Manab^{1,2}

¹School of Law and Government, Dublin City University, Dublin, Ireland

²School of Data and Sciences, BRAC University, Dhaka, Bangladesh

¹meem.manab2@mail.dcu.ie

²meem.arafat@bracu.ac.bd

ABSTRACT

As we keep rapidly advancing toward an era where artificial intelligence is a constant and normative experience for most of us, we must also be aware of what this vision and this progress entail. By first approximating neural connections and activities in computer circuits and then creating more and more sophisticated versions of this crude approximation, we are now facing an age to come where modern deep learning-based artificial intelligence systems can rightly be called thinking machines, and they are sometimes even lauded for their emergent behavior and black-box approaches. But as we create more powerful electronic brains, with billions of neural connections and parameters, can we guarantee that these mammoths built of artificial neurons will be able to forget the data that we store in them? If they are at some level like a brain, can the right to be forgotten still be protected while dealing with these AIs? The essential gap between machine learning and the RTBF is explored in this article, with a premonition of far-reaching conclusions if the gap is not bridged or reconciled any time soon. The core argument is that deep learning models, due to their structure and size, cannot be expected to forget or delete a data as it would be expected from a tabular database, and they should be treated more like a mechanical brain, albeit still in development.

Keywords: *Privacy, Data Protection, Right to be Forgotten, Machine Unlearning*

1 Preamble: The AI Boom

In the 1990s, AI research had come to all but a halt, with the failure of early LISP machines¹, the budget cut from the US Department of Defence's DARPA-funded Strategic Computing Initiative Program², and the frustratingly low performance of the rule-based expert systems³ being just some of the sore points of that epoch. Fast forward three decades, and we are in the middle of an AI Boom, with artificial intelligence-enabled tools having become as ubiquitous as mobile phones and as useful as GPS, if not more. If anybody doubts how far we have come in the last twenty years alone, comparing Clippy from Microsoft Office circa 2000⁴ with Apple's Siri, and then comparing Siri with today's ChatGPT should suffice very well.

A large motivation for this progress in AI was the near-simultaneous availability of large datasets and mass-produced GPUs. Datasets like ImageNet⁵, which has more than 14 million images annotated and hierarchially arranged by human participants, made it possible to test and refine AI models, and the incredible performance in 2012 of AlexNet⁶, a convolutional neural network model, to classify ImageNet images could be considered one of the watershed moments in recent AI history. It outperformed its closest competitor by 10.8%, with an error margin of only 15.3% when allowed to predict five options instead of one while we consider the model correct as long as the correct response is among these top five (this is the reported top-5 accuracy). This error rate had been unimaginable up to that point. But what is more interesting is that one of the minds behind AlexNet was Geoffrey Hinton, who originally developed with two other colleagues the mathematics behind multi-layered neural networks⁷ in 1986. Truth is, the mathematics and the ideas were all there for decades (since at least the 1940s,

according to Jürgen Schmidhuber⁸, another progenitor of AI, who recalls Soviet mathematicians Vladimir Vapnik and Andrey Kolmogorov as indispensable precursors), but it was only when GPUs, i.e. graphical processing units, became readily available in the market that practical neural networks could be built. This was in no small part thanks to the video gaming industry which led to ever cheaper versions of GPUs, and what has followed is a quasi-exponential rise in share prices of graphics chip manufacturers like Nvidia as they diversified into AI⁹. Why did neural networks require GPUs in the first place? At some level, these models, just like biological brains, are built around the idea of parallel processing¹⁰, and CPUs cannot do parallel computations effectively, while GPUs, or in common parlance, graphics cards, are built to do precisely this because they were originally designed to render high-resolution video games in real-time on graphical displays, which is inherently a parallel process.

2 The Legal Backdrop

Now, what does any of this have to do with the Right to be Forgotten in particular? The Right to be Forgotten, in the case of the European Union, is actually of a rather limited scope. Often referred to as the Right to Erasure (of Data) as enshrined in GDPR's Articles 17 and 19 and also Recital 65¹¹, it awards EU citizens the right to request the deletion of their personal data without any unreasonable delay. This can lead to the deletion of their records, for example, from search engine results, as immortalized in the European Court of Justice case "Google Spain v AEPD and Mario Costeja González" and its historic judgment from 2014¹². There have been talks of how this deletion may not be as instantaneous as we want it to be, due to, for example, search engines keeping large amounts of data in multiple servers and so-called cache-s (pronounced like cash) to make the search process as fast and smooth as possible, as Fosch-Villaronga et al. pointed out in their seminal work *Humans Forget, Machines Remember: Artificial Intelligence and the Right to Be Forgotten*¹³. But ultimately, this line of reasoning still sees data as information kept in some table or some database, and with machine learning, that might not quite be the case.

3 Welcome to the Machine

We should stress here that most AI technologies that have been either changing the world or making the headlines in the last decade, be they the facial recognition systems, deep fakes, or large language models like GPT-3 and Bard, are examples of machine learning, or more specifically its sub-branch called deep learning, where large amount of data is fed into an algorithm of layers of neural networks which identifies patterns in the existing data, and then either categorises or generates data using these patterns. These deep learning models employ internal parameters, known as weights and biases, for their tasks of pattern recognition and data generation. The parameter values are derived from some 'training' data provided to the algorithm and then used in the mathematical calculations for classification or creation of new data. Statistical in nature, the size of these models, i.e. the number of their parameters, has generally been smaller than the data they are trained with. Take GPT-4, as a reference. According to some sources, it has 1.7 trillion parameters and it was trained on 13 trillion tokens (small chunks of words and sentences)¹⁴, while GPT-3 has 175 billion parameters¹⁵. The information of these tokens is then remembered using the parameter values, so each token might be remembered using a handful of parameters, and each parameter could also be responsible for remembering some different tokens. Parameters, in this case, are doing what neurons do in our brains. If we want GPT-4 to "forget" the sentence: "Harry Potter can be treated as a fan-fiction of Lord of the Rings", assuming each word in this sentence is a token, deleting one or two parameter values from the 1.7 trillion should be enough.

The problem is that we do not know which parameter in reality captures which information. Some scholars have rightly utilized neural networks for encryption, rather than compression¹⁶. Even if a machine learning model is not smaller than the data it has been trained on, as is often the case for the likes of zero-shot¹⁷ and few-shot learning¹⁸ models, it is still quite impossible to reverse-engineer the model and say with certainty that these "neurons" or these parameters are responsible for remembering this specific piece of data and this piece only. Just like an animal brain, trying to delete one piece of data might lead to a loss of information related to something else. We may not even know with great certainty which records would thus be affected since we do not know which pieces of

information are correlated in the internal representation of the model. There have been recorded instances that deleting the information of one person can result in the deletion of information of other people with the same or similar names¹⁹. Researchers have been comparing neural networks to a hypothetical black box²⁰ since their early days²¹, and if these networks are not simply a collection of data, but rather a black box that consumes data and performs intelligent tasks based on them, deleting the data should be very difficult.

And it indeed is. Most machine unlearning research (unlearning is the process by which a machine learning model forgets its training data) shows an accuracy far below²² what would be acceptable on the legal grounds of protecting fundamental rights. Turning back a century, cases like *Melvin v. Reid* (1931) and *Sidis v. FR Publishing Corp.* (1940)²³ from the United States first brought legal attention to the problem of being forgotten from the public eye. It was in relation to having personal information published without consent in films (*Reid*), books, or newspapers (*Sidis*) that these plaintiffs sought justice. Questions about freedom of speech or freedom of expression or perhaps undue censorship can then be brought against this right to claim relative obscurity. As long as things were confined to tabular data stored in some Spreadsheet file and then being Googled by strangers, the parallels had been there. But when we started to upload data into digital or electronic brains, things became convincingly more difficult. How do you make sure that ChatGPT will not remember you? What if the machine starts to lie? It has already been documented that large language models can have their memory distorted when tasked with learning new data²⁴, or may even appear to hallucinate²⁵. Faced with the challenge to forget some data, the distortion in their memory can be catastrophic.

4 Path to Reconciliation

There is a rather obvious way toward resolving this conundrum, which is re-training the machine learning models from scratch, known in the literature as exact unlearning²⁶. The trade-off here is even more disproportionate, as sources believe that it took 34 days for 1024 servers²⁷ (that is 1024 high-quality CPUs and GPUs) to train GPT-4. Now, if we wanted the personal data of the authors of this article, for example, to be deleted from GPT-4, retraining it would, in addition to time and server, require around 50 GWh of energy²⁸, which is enough to power 44 million households in the US for one year (that is one-third of the United States, and 24 times the size of Ireland). The other option, to use the currently available machine unlearning algorithms, would mean deleting it with a non-zero error margin²⁹, which means that there will always remain a tiny but real chance that our data might not be deleted at all after our request has been processed. There is also a lot of ongoing research into differential privacy, which essentially means adding noise to all personal data for maximum anonymization. But it is still in an ongoing phase, and no large scale industrial development has been deployed yet.

What if neural networks stopped being black boxes? If we could look at a machine learning model and tell which data it was trained on, and which parameters correspond to which pieces of data, then deleting corresponding parameters would be easy and straightforward, but a lot of information would then become compromised. For example, Meta built its own large language model LLaMA³⁰ in 2023, and its weight parameters (in four versions, from 7 to 65 billion) have been freely available to download. If the data could be derived backward from these parameter values, we would have access to an abundance of text data that may or may not include people's personal data as well.

There are also issues previously discussed in other publications, such as, for example, how these machine learning models, especially the large language models, store their data in intermediary servers, again for the sake of speed. This approach, usually referred to as "lazy evaluation"³¹ in computer science literature, can lead to retaining some data although it has been deleted officially. But even without lazy evaluation and caching, machine learning and the right to be forgotten are fundamentally at odds with each other. With machine learning, we are ever more approaching the creation of an artificial brain, often panderingly called AGI (Artificial General Intelligence), and the question now morphs into this: should we be able to delete data from a brain? What if it affects the workings of the rest of that brain? As we moved from rule-based to probabilistic artificial intelligence, we also moved from tabular databases that looked like gigantic Spreadsheets to neural networks, and deletion would now mean more than selecting cells in a table and then deleting the corresponding records.

5 Coda: The Digital Amnesiac

None of the versions of the AI act³² currently available on the internet address this gap between machine learning and the right to erasure, although they go to a good length to explain what machine learning is, and some previous legal literature had sparingly acknowledged the disparity³³ before. Meanwhile, at least some computational researchers have argued that machine unlearning is only achievable mathematically, not practically³⁴. Several potentially prohibitive solutions are still available for implementation, such as barring all training of machine learning, essentially deep learning, models using biometric or personal data until a foolproof and guaranteed machine unlearning method is proven to exist for that particular model. Another way could be to make AI tools like ChatGPT respond that they are unauthorized to provide a particular individual's personal data, rather than delete this data from within the underlying model itself. But then arises the risk of somebody deceiving the model (be it GPT-4 or Bard) to divulge the information³⁵. Or perhaps, the European Court of Justice could redefine or explain further what the Right to be Forgotten should mean in the age of AI. Geoffrey Hinton, arguably the father of neural networks and whom we mentioned at the beginning of this article, prescribed in 2022 to create specific-purpose computers for next-generation AI models³⁶, effectively barring people from doing AI research at home and turning transfer of parameters meaningless, through what he calls "mortal computation". But while that would eliminate the risk of transferring data via transferring parameter, what will happen to the question of deletion? The authors of this article believe that as machine learning models become more and more like an anatomical brain, both in terms of performance and structure, the more irreconcilable they will become with the Right to be Forgotten. Can we ask a brain to forget us, even if we created it ourselves? It would perhaps be wiser, instead of taking reconciliatory attempts, to never introduce ourselves to that brain in the first place. Otherwise, in addition to forgetting one person named Clementine K.³⁷, the mechanical mind³⁸ could end up forgetting songs and fruits by the same name, with a digital amnesiac instead of an artificial general intelligence standing before us.

Data Availability

No data has been generated or analyzed during this study at any phase.

References

1. Withington, P. T. The lisp machine: Noble experiment or fabulous failure? *Proc. 40th Anniv. Lisp – Lisp In The Mainstream – Conf.* (1998).
2. Roland, A. & Shiman, P. *Strategic computing: DARPA and the quest for machine intelligence, 1983-1993* (MIT Press, 2002).
3. Harmon, P. The expert systems business: How it grew and died. *IEEE Annals Hist. Comput.* **44**, 31–43 (2022). DOI 10.1109/MAHC.2022.3149465.
4. Maedche, A., Morana, S., Schacht, S., Werth, D. & Krumeich, J. Advanced user assistance systems. *Bus. Inf. Syst. Eng.* **58**, 367–370 (2016).
5. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009). DOI 10.1109/CVPR.2009.5206848.
6. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12*, 1097–1105 (2012). DOI 10.5555/2999134.2999257.
7. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nat.* **323**, 533–536 (1986). DOI <https://doi.org/10.1038/323533a0>.
8. Schmidhuber, J. Annotated History of Modern AI and Deep Learning. *arXiv preprint arXiv:2212.11279* (2022). DOI 10.48550/arXiv.2212.11279.

9. Witt, S. How Jensen Huang's Nvidia is powering the A.I. revolution. <https://www.newyorker.com/magazine/2023/12/04/how-jensen-huang-s-nvidia-is-powering> (2023). Accessed: 2024-2-14.
10. Chetlur, S. *et al.* cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759* (2014). DOI 10.48550/arXiv.1410.0759.
11. European Parliament & Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). URL <https://data.europa.eu/eli/reg/2016/679/oj>.
12. Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González ECLI:EU:C:2014:317. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0131>.
13. Villaronga, E. F., Kieseberg, P. & Li, T. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Comput. Law & Secur. Rev.* **34**, 304–313 (2018). DOI 10.1016/j.clsr.2017.08.007.
14. Koubaa, A. GPT-4 vs. GPT-3.5: A concise showdown. *Prepr.* (2023). DOI 10.20944/preprints202303.0422.v1.
15. Kalyan, K. S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Nat. Lang. Process. J.* **6**, 100048 (2024). DOI 10.1016/j.nlp.2023.100048.
16. Patel, S. *et al.* Colour image encryption based on customized neural network and DNA encoding. *Neural Comput. Appl.* **33**, 14533–14550 (2021). DOI 10.1007/s00521-021-06096-2.
17. Xian, Y., Schiele, B. & Akata, Z. Zero-shot learning - the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). DOI 10.1109/CVPR.2017.328.
18. Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **53** (2020). DOI 10.1145/3386252.
19. Xu, J., Wu, Z., Wang, C. & Jia, X. Machine unlearning: Solutions and challenges. *arXiv preprint arXiv:2308.07061* (2023). DOI 10.48550/arXiv.2308.07061.
20. Savage, N. Breaking into the black box of artificial intelligence. *Nat.* (2022). DOI 10.1038/d41586-022-00858-1.
21. Castelvechi, D. Can we open the black box of AI? *Nat.* **538**, 20–23 (2016). DOI 10.1038/538020a.
22. Xu, H., Zhu, T., Zhang, L., Zhou, W. & Yu, P. S. Machine unlearning: A survey. *ACM Comput. Surv.* **56** (2023). DOI 10.1145/3603620.
23. Bates, S. The prostitute, the prodigy, and the private past. *Commun. Law Policy* **17**, 175–233 (2012).
24. Ye, H., Liu, T., Zhang, A., Hua, W. & Jia, W. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794* (2023). DOI 10.48550/arXiv.2309.06794.
25. Zhang, Y. *et al.* Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023). DOI 10.48550/arXiv.2309.01219.
26. Chundawat, V. S., Tarun, A. K., Mandal, M. & Kankanhalli, M. Zero-shot machine unlearning. *IEEE Transactions on Inf. Forensics Secur.* (2023). DOI 10.1109/TIFS.2023.3265506.
27. Lai, C. AI is harming our planet: addressing AI's staggering energy cost (2023 update) (2023). URL <https://www.numenta.com/blog/2023/08/10/ai-is-harming-our-planet-2023/>. Accessed: 2024-2-14.
28. Generative AI does not run on thin air! URL <https://www.ri.se/en/news/blog/generative-ai-does-not-run-on-thin-air>. Accessed: 2024-2-26.

29. Sekhari, A., Acharya, J., Kamath, G. & Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. *Adv. Neural Inf. Process. Syst.* **34**, 18075–18086 (2021). DOI 10.48550/arXiv.2103.03279.
30. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023). DOI 10.48550/arXiv.2302.13971.
31. Vanhoucke, V., Senior, A. & Mao, M. Z. Improving the speed of neural networks on cpus. In *Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011* (2011).
32. Madiega, T. Artificial intelligence act. *Eur. Parliam. Eur. Parliam. Res. Serv.* (2021).
33. Weissinger, L. Ai, complexity, and regulation. *OUP Handb. on AI Governance, Forthcom.* (2021). DOI <https://doi.org/10.1093/oxfordhb/9780197579329.013.66>.
34. Thudi, A., Jia, H., Shumailov, I. & Papernot, N. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*, 4007–4022 (2022).
35. Floridi, L. Machine unlearning: its nature, scope, and importance for a “delete culture”. *Philos. & Technol.* **36**, 42 (2023). DOI 10.1007/s13347-023-00644-5.
36. Hinton, G. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345* (2022). DOI 10.48550/arXiv.2212.13345.
37. Gondry, M., Kaufman, C. & Brion, J. *Eternal sunshine of the spotless mind* (Universal Universal City, 2004).
38. Crane, T. *The mechanical mind: A philosophical introduction to minds, machines and mental representation* (Routledge, 2015).