EMBRACING AI

S. D. ROBINSON

Fraign Analytics LLC
222 S Meramec
Suite 202
St. Louis, MO
63105-3504

Abstract. Artificial intelligence's many forms have one thing in common—accuracy optimization. Although a seemingly sensible concept, it differs drastically from the human process of situated learning. This, at times, subtle difference can lead to potentially dangerous results that are couched in innocuous terms like "hallucinations." Rather than being something that can be resolved by model improvement, these hallucinations are examples of features in data that are fragile and incomprehensible to humans, but no less legitimate. A different approach is required if we are to reap the very real benefits to productivity that artificial intelligence offers. By considering artificial intelligence as part of a distributed system in cooperative partnership with humans, we can mitigate its weaknesses and leverage our human strengths. With this focus, we are able to craft holistic solutions, gain individual and public trust, as well as comply with the law. People are exceptional at solving novel problems; we need to ask ourselves how we can embrace artificial intelligence to make our jobs and lives better.

A primer on the history of artificial intelligence

1943 Perceptron neural network was introduced.

The history of artificial intelligence (AI) is surprisingly long, beginning in the 1940s with the *perceptron*. This was an analog system with inputs from a 400-pixel camera. From the 1940s until the 2000s, there were significant advances in the field in all aspects of machine learning and neural networks. Applications included voice transcription, natural language processing—that includes machine translation, text summarization—machine vision, handwriting recognition, and genetic algorithms. However, in the more recent past, a paradigm shift came with the use of graphics cards for model training of neural networks.

2004 20x training speed increase with GPUs.

Oh and Jung (2004) made a key insight into the nature of neural network training and consumer graphics processing units (GPUs) that quickly became the dominant approach (Strigl et al., 2010; Ciresan et al., 2011; Schmidhuber, 2015). GPUs are optimized to render images quickly by dividing complex mathematical operations into smaller ones that can be distributed across many tiny processing units. Oh and Jung's break-through was to recognize that mathematically, this is a nearly identical processes to that necessary for training inference in neural networks. This insight brought about a twenty-fold improvement in performance with the graphics cards of the day. Today the difference between a high-end CPU and GPU is even greater. Considering that large language models (LLM) like ChatGPT require months of training time, without the use of GPUs, it could have taken close to a decade.

2017 Transformer attention model published.

The "T" in ChatGPT is from the transformer model architecture that was introduced by a Google research team in December 2017 in a paper titled "Attention is all you need" (Vaswani et al., 2017). As is frequently the case in revolutionary approaches, it was published without fanfare simply as a higher-performance, more cost-effective model for language translation tasks. Following its publication, all large language models have moved to this architecture.

A complete LLM requires more than 10GB of GPU memory per billion parameters.

The transformer is a generative model in that it is able to predict the next word in a sequence and generate new tokens given a set of prior words. In this approach, the importance and position of prior words are encoded numerically in large matrices; this is called an "attention mechanism." Unlike prior approaches, which had the tendency to "forget" long sequences, the memory of the transformer's attention mechanism scales with the size of the matrices. Not only does this approach allow for greater memory, but it also allows for parallel computation of those matrices. The result is that transformers are both more capable and also faster to compute. The size of that memory is limited and of fixed length. The memory of ChatGPT is impressive at thirty-two thousand tokens, but it's still a proverbial "goldfish" (Stern, 2023; Wiggers, 2023).

Date: June 21, 2023.

AI governance and regulation

A question for any user of ChatGPT is, "Have you read the licensing agreement?" If you have not, there may be significant ramifications for you and your business. Navigating legal compliance and maintaining intellectual property while being effective in your mission is difficult enough for any organization. Big tech has the intent of securing their dominance with LLMs—often described as creating a moat around their market. But transforming research into a commercial product isn't so simple, especially when a major competitor makes their model openly available (Meta AI, 2023) and you don't own the copyright on the materials that were used to train your models.

LLM training data was scraped from the internet and is broadly copyrighted.

The data being used to train the current batch of large language models is not owned by OpenAI, Meta, Google, or any one company. That textual data was scraped from the internet. The result has been lawsuits citing a range of violations including the Digital Millennium Copyright Act (Claburn, 2023). Not only is there public work within the training data, but there is almost certainly illegally obtained copyrighted material within that data as well (Chris, 2023). It's known that within the Common-Crawl dataset—which all of the large language models leverage—there are hundreds of thousands of instances of the copyright symbol alone. Everyone's internet data is there and it was likely used to train one of these models.

Copyright is one aspect of legality, but other criminal laws cover developments with large language models. The first and perhaps most dangerous for AI companies is data protection (Weatherbed, 2023). These laws protect individual citizens' right to privacy and their well-being. OpenAI has been opaque with their model specifications. However, we know that Meta AI (2023) model was trained on over one point four trillion tokens. One question that arises is how can a model vendor comply with data protection's "right to be forgotten" (European Commission, 2023)? Removing a person's data from an LLM is a problem with no readily available solution. First the offending data must be removed from the unstructured dataset. In and of itself, defining these searches and then parsing terabytes of data has a massive human and computational overhead. Then given a compliant dataset, the model must be retrained from scratch. While obstacles remain in the commercialization of LLMs, academic and public research continues unabated.

There is no simple way for an LLM to comply with data protection laws.

Meta's LLaMa model is openly available.

Following Meta's release of their 65 billion-parameter LLaMA model to academia—and its subsequent leak to the public domain—LLMs are now widely available (HuggingFace, 2023). Although complete model training is not feasible without specialized hardware, "parameter efficient" methods make it possible to fine-tune LLMs without major investment (Lialin et al., 2023; Houlsby et al., 2019; Ziegler et al., 2019). One of the most effective of these optimizations was introduced by Microsoft Research in 2021 (Hu et al., 2021). With this approach there is no need to retrain all of the parameters of the model, just a small fraction. This method allows for as much as a ten thousand-fold reduction in the parameters needed for training and a three-fold reduction in memory requirements. As a result, "citizen science" is possible and the field is accelerating exponentially. No where is this more apparent than in a leaked memo—allegedly from Google—stating that "[t]he barrier to entry for training and experimentation has dropped from the total output of a major research organization to one person, an evening, and a beefy laptop"(Patel and Ahmad, 2023).

Although they are making substantial progress toward AI regulation (European Commision, 2020; EASA, 2021; Shepardson and Diane, 2023; Panchanathan and Prabhakar, 2023), given the rate of development, regulators acknowledge that they have a difficult road ahead (Gavaghan et al., 2019; Meltzer, 2023; Shepardson and Diane, 2023; Bennet and Welch, 2023).

Security

When addressing security in software, the traditional focus is on protecting applications from malicious actors. But with AI, we also have to consider "hallucinations" and additional serving costs as vulnerabilities. The hallucinations of large language models fall into the broader category of "adversarial attacks". These adversarial attacks are an intrinsic part of AI (Tanay and Griffin, 2016; Schmidt et al., 2018; Gilmer et al., 2018; Fawzi et al., 2018; Shafahi et al., 2018).

Mathematically, AI models maximize accuracy above all else. As a result, they exploit the most effective predictive features from the data, whatever they may be. That data may contain features that, when viewed from a human-selected notion of similarity, are fragile. For an image classification

Adversarial attacks and hallucinations are fundamentally human phenomena.

task, this might mean that a picture of a bus is labeled as an ostrich. Which, from the human perspective, is comical, yet it demonstrates that images contain features that we simply aren't aware of as humans.

These adversarial attacks have tremendous ramifications for both security and governance, albeit subtle ones. As the research of Ilyas et al. (2019) demonstrates:

> Adversarial examples can be directly attributed to the presence of non-robust features: features derived from patterns in the data distribution that are highly predictive, yet brittle and incomprehensible to humans. After capturing these features within a theoretical framework, we establish their widespread existence in standard datasets.

To restate a major point of Ilyas et al.'s work, these brittle incomprehensible features are present in all datasets. In addition to this new consideration, we must also acknowledge the increased costs of AI.

AI services must have engineered resilience to abuse, bad actors, and internal errors, just as any service must have. However, the operational costs for large AI models such as ChatGPT are orders of magnitude higher than non-AI services. Thus, the resilience of a service transitions from an annoyance to a potential business disruptive cost. In order to avoid such costs without adversely affecting the user and developer experience, it is essential for organizations to invest in additional orchestration tooling.

AI serving costs are high.

Adoption & ownership

AI adoption requires a thorough understanding of the front-line workers' daily issues and leadership's identified priorities. Once the organization's problems are laid bare, it requires subject matter expert involvement on all sides. Knowing where AI can be applied and how effective the implementation can be is a trade-off that can only be evaluated as part of a concerted effort that is driven by example (Adzic, 2011; Kim et al., 2023). That trade-off requires an acute understanding of the highly abstract limitations of AI.

Algorithms do not replace expertise.

Ownership and buy-in often mature when models are run in a "shadow deployment" first (Kim et al., 2023). Not only does this approach ensure that the experience of the end-user is what was expected, but it allows for a feedback loop for continuous improvement. Since users are engaged in the development process—making their opinion subjectively and objectively matter—buy-in improves. This incremental approach has the additional benefit of mitigating risk. Through this process of integration, the risk is limited first by the scope and then by review prior to broader deployment. Since the review process engages everyone involved in the workflow, not just data scientists and managers, the organizational risk is minimized.

Distributed cognition & cooperative AI

AI is being built to complete tasks in autonomous ways (Walch, 2020). Yet most productive organizations have humans and machines working together (Wilson and Daugherty, 2018). In order to excel with this new technology, organizations need to leverage the strengths of AI and the humans involved.

The process begins with understanding both participants in situ. This understanding requires a process that requires subject matter experts on both sides. It then requires that the AI is trained to complete the task fragment in the manner that makes the most sense to the humans involved. In this way, we can accelerate our productivity and also our ability to make effective decisions.

In order to understand how to create an effective cooperative relationship between AI and humans, we need to first understand how cognition can be divided among multiple agents and their environment. In the case of aviation, a cooperator has been part of the cockpit for many decades—the autopilot.

Aviators are already familiar with cooperative AI—the autopilot.

The autopilot serves as a component in the aircraft system, meshing flight crew and machine (Hoeft et al., 2006). From initial flight training through the rest of their careers, pilots are trained to appreciate the limitations and practical usage of the autopilot available in their aircraft (Curry, 1985; Sarter and Woods, 2017). The division of duties that the autopilot provides reduces the pilot's workload during normal flights. However, without a vigilant continuous focus on the outcomes of the mission by the pilot, in a state termed "situational awareness", the autopilot can be detrimental to

the overall safety of the operation. In this way, when asked "Who is flying the airplane?", the pilot's response should never be "the autopilot" (Nutter and Anthony, 2020), as the autopilot will happily drive the aircraft into an unsurvivable situation. Similarly, when asked "Who is driving the project?", the answer is never "the AI."

Conclusion

AI is a perfect student, one that doesn't learn the really human practical lessons. The hallucinations or adversarial examples are features intrinsic to the data; they are not bugs to be resolved by an improved model. If there is to be compliance with civil law and regulation, as well as public trust, data must be provably fit for purpose. Although AI has its limits, it has tremendous strengths that, by considering it in a cooperative and distributed way, we can leverage. Technical development is a process of solving novel problems, and this is something that humans are very good at.

References

Adzic, G. (2011). *Specification by example: how successful teams deliver the right software*. Simon and Schuster.

Bennet, M. and Welch, P. (2023). Bennet, Welch reintroduce landmark legislation to establish federal commission to oversee digital platforms. `https://www.bennet.senate.gov/public/index.cfm/`.

Chris, S. (2023). ChatGPT seems to be trained on copyrighted books like Harry Potter. *New Scientist*.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence*. Citeseer.

Claburn, T. (2023). GitHub, Microsoft, OpenAI fail to wriggle out of Copilot copyright lawsuit. *The Register*.

Curry, R. E. (1985). The introduction of new cockpit technology: A human factors study. Technical report, National Aeronautics and Space Administration.

EASA (2021). EASA concept paper: First usable guidance for level 1 machine learning applications. `https://www.easa.europa.eu/en/first-usable-guidance-level-1-machine-learning-applications-issue-01`.

European Commision (2020). Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment. `https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment`.

European Commission (2023). Everything you need to know about the "right to be forgotten". `https://gdpr.eu/right-to-be-forgotten/`.

Fawzi, A., Fawzi, H., and Fawzi, O. (2018). Adversarial vulnerability for any classifier. *Advances in neural information processing systems*, 31.

Gavaghan, C., Knott, A., MacLaurin, J., Zerilli, J., and Liddicoat, J. (2019). Government use of artificial intelligence in new zealand. `https://www.data.govt.nz/assets/data-ethics/algorithm/NZLF-report.pdf`.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. (2018). Adversarial spheres. *arXiv preprint arXiv:1801.02774*.

Hoeft, R. M., Kochan, J. A., and Jentsch, F. (2006). Automated systems in the cockpit: Is the autopilot,"george," a team member? *American Psychological Association*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

HuggingFace (2023). HuggingFace Models. `https://huggingface.co/models`.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

Kim, J. Y., Boag, W., Gulamali, F., Hasan, A., Hogg, H. D. J., Lifson, M., Mulligan, D., Patel, M., Raji, I. D., Sehgal, A., et al. (2023). Organizational governance of emerging technologies: AI

adoption in healthcare. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1396–1417.

Lialin, V., Deshpande, V., and Rumshisky, A. (2023). Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*.

Meltzer, J. P. (2023). The US government should regulate AI if it wants to lead on international AI governance. *Brookings*.

Meta AI (2023). Introducing LLaMA: A foundational, 65-billion-parameter large language model. `https://ai.facebook.com/blog/large-language-model-llama-meta-ai/`.

Nutter, C. and Anthony, T. (2020). Who has the airplane? `https://flightsafety.org/asw-article/who-has-the-airplane/`.

Oh, K.-S. and Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314.

Panchanathan, S. and Prabhakar, A. (2023). Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: An implementation plan for a national artificial intelligence research resource. `https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf`.

Patel, D. and Ahmad, A. (2023). Google "We have no moat, and neither does OpenAI". `https://www.semianalysis.com/p/google-we-have-no-moat-and-neither`.

Sarter, N. B. and Woods, D. D. (2017). Pilot interaction with cockpit automation ii: An experimental study of pilots' model and awareness of the flight management system. In *Situational Awareness*, pages 259–286. Routledge.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. (2018). Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. (2018). Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*.

Shepardson, D. and Diane, B. (2023). US begins study of possible rules to regulate AI like ChatGPT. *Reuters*.

Stern, J. (2023). GPT-4 has the memory of a goldfish. *The Atlantic*.

Strigl, D., Kofler, K., and Podlipnig, S. (2010). Performance and scalability of GPU-based convolutional neural networks. In *2010 18th Euromicro conference on parallel, distributed and network-based processing*, pages 317–324. IEEE.

Tanay, T. and Griffin, L. (2016). A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Walch, K. (2020). The autonomous systems pattern of AI. *Forbes*.

Weatherbed, J. (2023). OpenAI's regulatory troubles are only just beginning. *The Verge*.

Wiggers, K. (2023). Antropic's latest model can take 'The Great Gadsby' as input. *TechCrunch*.

Wilson, H. J. and Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Fraign Analytics LLC, 222 S Meramec, Suite 202, St. Louis, MO, 63105-3504
Email address, S. D. Robinson: `saul@fraignanalytics.com`