Protocol Learning, Decentralized Frontier Risk and the No-Off Problem

Alexander Long
Pluralis Research
alexander@pluralis.ai

Abstract

Frontier models today are either trained centrally and available behind paid API's, or trained centrally and opensourced. There appears to be the possibility of a third approach; Protocol Learning, where models are sharded across nodes and trained within an elastic pool of independently controlled compute consisting of multiple participants. This setting comes with significant technical challenges, however if instantiated would significantly alter the landscape of frontier model risk due to both novel the governance structures introduced and potentially unprecedented scale. To date, there has been no analysis of either the feasibility of such an approach or the risks such an approach would introduce. We summarize the prior art and conclude Protocol Learning may be significantly more feasible than researchers are currently aware. As decentralization circumvents centralized governance efforts, we extensively discuss the risks associated and argue that Protocol Learning reduces rather than increases frontier risk.

1 Introduction

Fundamental modelling advances [39] combined with unprecedented scale have resulted in models able to perform routine knowledge work to a level beyond the standard human [1, 6, 8]. Current trends give every indication that increasing scale will continue to increase model performance [15, 28]. Calls for pauses to this line of research have, to date, been entirely unsuccessful [11]. While continued increase in model capability seems likely, it is not clear at which point such models gain the ability to solve certain tasks [33, 40], or if they will at all. Consequently, despite widespread understanding of the misalignment and misuse risks such models pose, because the front of tasks being solved has clear economic utility there is an ongoing race to train larger models at larger and larger scale. This is occurring at enormous cost, with a corresponding expectation of enormous value capture.

A prevailing view is opensource models will provide a counterbalance to the emerging frontier model oligopoly. This ignores the huge cost of training such models. Releasing the output of a process that requires hundreds of millions to billions of dollars of cost for free, without constraints, is unsustainable and will not continue. There is hence clear motivation for decentralized training as it would reduce the opensource movements dependency on centralized training runs.

2 Decentralized, Centralized and Volunteer Network Computational Capacity

The scale of centrally controlled compute capacity is perhaps underappreciated; it is enormous today and growing rapidly. Precise figures are not publicly disclosed however as a single public datapoint, Meta has announced plans to purchase 350k H100s by year end 2024 [38] - on the order of 350 exaFLOPS at theoretical peak load using TF32 datatype with sparsity feature [27].

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

In contrast, the maximum compute capacity achieved by volunteer networks was a temporary peak of 1.2 full-precision exaFLOPS by the Folding at Home Project [20] in March 2020. Over 2 million devices (1.4M of which were CPU's) were present in the swarm, triggered by a surge of interest in projects simulating theoretically druggable protein targets from SARS-CoV-2. In short; volunteer network capacity peaked at two orders of magnitude below a single centralized actors compute purchases in a single year.

Compared to volunteer networks and centralized clusters, incentivized decentralized swarms, such as those assembled for Proof-of-Work (PoW) mining in the Bitcoin [25] and Ethereum [42] protocols have achieved orders of magnitude larger capacity than any centralized cluster. We measure productive capacity here in terms of Watts rather than FLOPS in order to make meaningful comparisons (PoW mining does not involve any floating point operations). 350k H100's running continually at peak power draws 0.24 GW (1 GW is the average energy consumption rate of a one-million inhabitant industrial city). In contrast; Bitcoin PoW mining consumption is estimated at 150 ± 50 TWh in 2022 [24], or 17.12 GW on average and approximately 0.5% of total worldwide energy consumption. While these figures are approximate the core fact remains striking; given certain incentives, pooled compute two orders of magnitude larger than the already enormous largest pools of centrally controlled compute can, and has, been assembled under a single protocol.

3 Feasibility of Protocol Learning

A common objection to large-scale decentralized training over the internet is that it is infeasible given the node-node communication speeds are orders of magnitude slower than those required for frontier model training in centralized clusters. Recent work has challenged this view and proposed approaches to low-bandwidth training with only minor overhead. The majority of early results are in the Distributed Data Parallel (DDP) setting [21], with full model replicas per node. Here, rather than communicating in a specific topology synchronously, nodes are allowed to drift, and communicate with neighbors via gossip protocols [3, 5] or similar. Convergence guarantees can still be obtained in this setting [22, 23] even with the communication graph altering during training [19, 35].

Decentralized training with heterogeneous devices and low bandwidth connections has also been demonstrated, but constrained to small models. Moshpit-SGD [31] introduces an approach that is both communication efficient and scales well with heterogeneous compute and communication bandwidth. Diskin et al. [9] perform a real run, over 200 MB/s interconnects, using devices with a range of capabilities on a dynamic swarm to train a 72.5M parameter ALBERT-xlarge variant. Ryabinin et al. [32] practically demonstrating the training of a 1B parameter LLM on T4 GPU's with 500 MB/s interconnects, achieving roughly 20% throughput overhead to centralized training, maintaining very high accelerator utilization, and also possessing basic fault tolerance. This is achieved with redundancy within each pipeline stage, dynamic routing between each stage, and the assumption of good actors. Learning@Home [30] propose the Decentralized MoE architecture and an asynchronous training scheme in order to achieve communication efficiency over heterogeneous nodes. Such an approach can theoretically scale to very large parameter sizes but has not been practically demonstrated beyond 257M, and while node failures are handled, byzantine nodes are not.

As noted in Sec. 2, decentralized training has the ability to assemble several orders of magnitude larger compute capacity than any centralized actor and consequently the ability to train models orders of magnitude larger than any today. It is our view that in order to achieve this scale (i.e. for the system to obtain the underlying compute and data required to train state of the art models), decentralized training must be directly incentivized. We term decentralized training, when combined with explicit incentives, *Protocol Learning*. There are many approaches to incentivization, however fractional ownership allocated proportional to the training contribution is appealing as it both directly aligns incentives and results in a self-contained system. As there is real cost to contributing (the cost of compute), trainers will not contribute to models that are not high expected utility (or if they do, it will be with the understanding they will make no return). This in turn creates market forces that push contributors towards development of the highest utility models for the lowest cost, not only in algorithm and learning design but also in hardware setup and location (due to energy prices, operating costs, etc.). However, such an approach also introduces novel risks.

4 Decentralized Frontier Risk

As noted in Sec. 2, incentivized protocols have the ability to assemble several orders of magnitude larger compute capacity than any centralized actor. It is our view that there are a range of approaches able to combine incentivization with the rapidly advancing progress in decentralized training, to facilitate training runs larger than current frontier models. This would result in a new class of models, which would alter the landscape of fronteir risk.

Standard (Centralized) Frontier Risk assumes a single actor trains, owns and distributes the model, and focuses on two main themes separated along an axis of base model capability; misalignment [17], and misuse [7]. Common misuse risk categories include; cybersecurity [4], persuasion [12] and chemical and biological threats [37], and are feasible with models today or are predicted to be feasible in the short term future. Misalignment risks largely revolve around more capable models, that have achieved some level of autonomy [41]. While there is a large emerging body of work on frontier risk to date no analysis of Decentralized Frontier Risk (DFR) has been carried out. DFR specially focuses on frontier models trained in a decentralized manner with decentralized governance. Implicit in this definition is that incentives are present; we argue in Sec. 2 incentives are required to reach the scale possible to train a frontier model in the decentralized setting. We discuss only portions of the AI risk landscape that are altered by decentralized models, and note many of the problems raised in the safety literature remain applicable to decentralized models.

4.1 Concentration of Power at the Organization Level

Frontier model value is becoming increasingly apparent. Such models are likely to form a base dependency for a large potion of software due to their improved generality and capability over all previous systems. Consider, for example, the 'AI tutor' scenario, where frontier models are adopted within most levels of education to assist students learning. This is already occurring [43] and appears to be a major short term application. While clearly beneficial over a textbook for queries such as 'can you explain integration by parts?' due to ability to be interactive, contextual and personalized, such systems would directly mold the worldview of students when answering questions such as 'how does an oligopoly form? Are oligopolies a bad thing?'. Differently to traditional learning where worldviews are shaped by various competing sources, opinions and individuals, if AI tutors become ubiquitous, and are powered by a small number of base models, worldviews would largely be shaped by a single source. Furthermore, in the current scenario, the design of this source would not be public (see Sec. 4.3).

4.2 Concentration of Power at the State Level, Geopolitical Risk and Rates of Progress

Polarization of capability will also occur at the state level. As models continue to grow and costs increase, it will not be possible for smaller developed nations to train their own frontier models (see Sec. 2), and as a consequence they will be relegated to model consumers, with only soft influence on their design and behaviour. This is a problem for such countries because ubiquitous frontier model use will likely have direct cultural impact. In the centralized case the only path to model ownership for such countries is for training costs to decrease and scaling to no longer result in meaningful model improvements. In contrast decentralized training potentially allows such governments, and the citizens of such governments to contribute their compute to various base models and participate directly in partial ownership and design decisions, without the need for massive capital outlay and infrastructure buildup required for standalone model development.

Consider also the impact of regulation when ability to train is highly skewed at the state level but use is not. Significant efforts are underway to regulate the capabilities made available in both centralized [10] and opensource [13] FM development. As FM development is currently centered solely within the largest countries, the governments and regulatory bodies of those countries supersede the control of the organizations creating the models, and hence can implement such regulation. This is perhaps reasonable [26] for the country in which the model was developed, however for model users in other countries (which will be the majority of users), such governance was not chosen.

4.3 Lack of Transparency and Poisoned Model Risk

In the current scenario, training recipes, data mixes, architectures and other design decisions are closely guarded trade secrets and are not released to model users. Users are able to evaluate the immediate obvious utility of the model but have no grantees around fine-grained behaviour in specific scenarios. It is likely then for use to accrue to the most broadly *accurate* and *useful* model, even if this model contains unacceptable or dangerous behaviour in specific areas injected either maliciously via model poisoning techniques [16], or are emergent and undetected. In the decentralized scenario, model design is public, the data mix is know, algorithms are known and any alterations are known, allowing informed analysis and use and removing this entire risk category.

4.4 The No Off Problem

We view this as the core risk introduced by Protocol Learning. The existential threat/rouge AI risk [14] scenario may seem speculative however it has received serious recent concern and study [2, 34]. In a centralized scenario, servers can be unpowered and datacenters can be isolated at the direction of a small number of individuals. In a decentralized scenario, as long as some portion of the swarm sufficient to support the model size remains online, the underlying model continues to operate. In a scenario where the model has achieved a level of performance able to influence human actors (note this does not require self-awareness or agency) and either alignment techniques have failed, trainers have specifically designed it to do so, or for some other reason, the model can continue to attract compute into the swarm. Consequently it is significantly more difficult to stop an unaligned AGI in this scenario; all participants must agree to pull compute from the network within a short time-frame.

Particular to Protocol Models is that this problem is exacerbated by any incentive structure where returns are tied to model performance; such a model will almost certainly be perceived to have high utility and hence, swarm participants will expect to receive large returns when contributing to training. All participants must then pass up on expected returns in order to stop model operation. Not only must existing trainers pull their contribution from the swarm, but new participants must also not join and take up the positive return that would be expected from contributing to a high utility model.

The magnitude of this problem is dependent of the form of work verification implemented within the protocol supporting incentivized decentralized training. If game-theoretic verification is implemented the no-off problem is reduced; if a large run is deemed by external actors to be dangerous, it would be possible to spend large amounts to derail the training run by repeatedly joining the run and contributing bad gradients. Such *model derailment attacks* would cost the attacker significantly (more than the cost incurred by legitimate trainers) and would gain them no economic utility other than preventing the run; however in some situations this could be seen as rational. However if exact work verification is developed with low overhead, such attacks no longer become possible and there are almost no ways for external actors to stop or pause a model running within the swarm aside from disrupting the run at the physical layer. We hence view the no-off problem as the most significant – albeit long term – risk of Protocol learning.[18, 29, 36].

If the game-theoretic verification is implemented the no-off problem is reduced; if a large run is deemed by external actors to be dangerous, it would be possible to spend large amounts to derail the training run by repeatedly joining the run and contributing bad gradients. Such *model derailment attacks* would cost the attacker significantly (more than the cost incurred by legitimate trainers) and would gain them no economic utility other than preventing the run; however in some situations this could be seen as rational. However if exact work verification is developed with low overhead, such attacks no longer become possible and there are almost no ways for external actors to stop or pause a model running within the swarm aside from disrupting the run at the physical layer.

5 Conclusion

Many of the required components for Protocol Learning have already been produced, but remain to be assembled within a single system and demonstrated at frontier model scale. When combined with explicit incentives, there is a realistic path towards training orders of magnitude larger models than exist today. Such systems would introduce novel risk categories and alter the overall risk landscape of frontier model development significantly. We argue such systems reduce Frontier Risk, and more fairly democratize Frontier Model access.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Yoshua Bengio. How Rogue AIs may Arise, 2023. URL `https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/`.

[3] Michael Blot, David Picard, Matthieu Cord, and Nicolas Thome. Gossip training for deep learning. *arXiv preprint arXiv:1611.09726*, 2016.

[4] Matteo E Bonfanti. Artificial intelligence and the offence-defence balance in cyber security. *Cyber Security: Socio-Technological Uncertainty and Political Fragmentation. London: Routledge*, pages 64–79, 2022.

[5] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Miles Brundage, Katie Mayer, Tyna Eloundou, Sandhini Agarwal, Steven Adler, Gretchen Krueger, Jan Leike, and Pamela Mishkin. Lessons learned on language model safety and misuse, 2022.

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[9] Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Anton Sinitsin, Dmitry Popov, Dmitry V Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, et al. Distributed deep learning in open collaborations. *Advances in Neural Information Processing Systems*, 34:7879–7897, 2021.

[10] FTC. FTC Launches Inquiry into Generative AI Investments and Partnerships, January 2024. URL `https://archive.md/3EzdQ`.

[11] Future of Life Institute. Pause Giant AI Experiments: An Open Letter, 2023. URL `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`.

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

[13] David Evan Harris. Open-Source AI Is Uniquely Dangerous - IEEE Spectrum, 2024. URL `https://spectrum.ieee.org/open-source-ai-2666932122`.

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[16] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[17] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[18] Sunny King and Scott Nadal. Ppcoin: Peer-to-peer crypto-currency with proof-of-stake. *self-published paper, August*, 19(1), 2012.

[19] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[20] Stefan M Larson, Christopher D Snow, Michael Shirts, and Vijay S Pande. Folding at home and genome at home: Using distributed computing to tackle previously intractable problems in computational biology. *arXiv preprint arXiv:0901.0866*, 2009.

[21] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.

[22] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

[23] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pages 3043–3052. PMLR, 2018.

[24] Irene Messina. Bitcoin electricity consumption: an improved assessment - News & insight, August 2023. URL `https://www.jbs.cam.ac.uk/2023/bitcoin-electricity-consumption/`.

[25] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*, 2008.

[26] David Nguyen, Valérie Frey, Santiago González, and Monica Brezzi. Survey design and technical documentation supporting the 2021 oecd survey on drivers of trust in government institutions. 2022.

[27] NVIDIA. NVIDIA H100 Tensor Core GPU Architecture Overview, 2023. URL `https://resources.nvidia.com/en-us-tensor-core`.

[28] David Owen. How predictable is language model benchmark performance?, 2024.

[29] Doug Petkanics and Eric Tang. Livepeer whitepaper. *Technical report, Livepeer*, 2018.

[30] Max Ryabinin and Anton Gusev. Towards crowdsourced training of large neural networks using decentralized mixture-of-experts. *Advances in Neural Information Processing Systems*, 33: 3659–3672, 2020.

[31] Max Ryabinin, Eduard Gorbunov, Vsevolod Plokhotnyuk, and Gennady Pekhimenko. Moshpit sgd: Communication-efficient decentralized training on heterogeneous unreliable devices. *Advances in Neural Information Processing Systems*, 34:18195–18211, 2021.

[32] Max Ryabinin, Tim Dettmers, Michael Diskin, and Alexander Borzunov. Swarm parallelism: Training large models can be surprisingly communication-efficient. *arXiv preprint arXiv:2301.11913*, 2023.

[33] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

[34] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. 2023.

[35] Zhenheng Tang, Shaohuai Shi, and Xiaowen Chu. Communication-efficient decentralized learning with sparsification and adaptive peer selection. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 1207–1208. IEEE, 2020.

[36] Jason Teutsch and Christian Reitwießner. A scalable verification solution for blockchains. In *ASPECTS OF COMPUTATION AND AUTOMATA THEORY WITH APPLICATIONS*, pages 377–424. World Scientific, 2024.

[37] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

[38] Jonathan Vanian. Mark Zuckerberg indicates Meta is spending billions of dollars on Nvidia AI chips, January 2024. URL `https://archive.md/UqV1x`.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[41] Norbert Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131 (3410):1355–1358, 1960.

[42] Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. 2014.

[43] Ismail Yesir and Danda B Rawat. Recent advances in artificial intelligence enabled tutoring systems: A survey. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0375–0381. IEEE, 2023.