The Future of Human-Artificial Intelligence Nexus and its Environmental Costs

Petr Spelda[1] and Vit Stritecky

Faculty of Social Sciences, Charles University

(Preprint, forthcoming in *Futures*[2])

ABSTRACT

The environmental costs and energy constraints have become emerging issues for the future development of Machine Learning (ML) and Artificial Intelligence (AI). So far, the discussion on environmental impacts of ML/AI lacks a perspective reaching beyond quantitative measurements of the energy-related research costs. Building on the foundations laid down by Schwartz et al., 2019 in the GreenAI initiative, our argument considers two interlinked phenomena, the gratuitous generalisation capability and the future where ML/AI performs the majority of quantifiable inductive inferences. The gratuitous generalisation capability refers to a discrepancy between the cognitive demands of a task to be accomplished and the performance (accuracy) of a used ML/AI model. If the latter exceeds the former because the model was optimised to achieve the best possible accuracy, it becomes inefficient and its operation harmful to the environment. The future dominated by the non-anthropic induction describes a use of ML/AI so all-pervasive that most of the inductive inferences become furnished by ML/AI generalisations. The paper argues that the present debate deserves an expansion connecting the environmental costs of research and ineffective ML/AI uses (the issue of gratuitous generalisation capability) with the (near) future marked by the all-pervasive Human-Artificial Intelligence Nexus.

[1] petr.spelda@fsv.cuni.cz

[2] https://doi.org/10.1016/j.futures.2020.102531

1. Introduction

Conceived as a scholarly discipline, ML seeks to develop ‘tools-for-optimal-action’. Given a task and evidence that can facilitate its mastering, the ‘tool-for-optimal-action’ earns its rank by being able to generalise. Such a tool then supports inferences which can generalise beyond the evidence (training data), i.e. can run inferences on new samples, provided that these come from the same, or a sufficiently similar, probability distribution as the evidence (training data). The discipline has a twofold epistemic aim. First, its theoretical purview, established by statistical learning theory, involves formal assumptions about the learning that leads to generalisations (Vapnik, 1995; Kawaguchi et al., 2019). Second, from the empirical viewpoint, the discipline seeks to improve the accuracy of inferences that furnish the acquired generalisations. At the moment, the cross-fertilisation between the two sub goals seems to be rather recalcitrant creating the following asymmetry. Although investing heavily in the theoretical research (e.g. cf. Zhang et al., 2017; Barlett et al., 2017; Neyshabur et al., 2017; Kawaguchi et al., 2019; Arjovsky et al., 2019), the field remains dominated by the second sub goal. As strong empirical results outpaced mature theoretical understanding, task-specific, accuracy-tracking leaderboards became the go-to measure for assessing the field’s epistemic progress.

The emphasis put on a single objective – accuracy – inspires a naïve idea that leaderboards are like ladders. The higher the rung, the closer we are to alleviating our cognitive burden by employing almost perfect ‘tools-for-optimal-action’ to carry out all sorts of tasks. If we construe the epistemic aim of ML as understanding generalisations, incentivising leaderboards produces troubles for the discipline itself and, quite strikingly, for the environment as well. The issue concerns the cost of computational resources that enable climbing to ever higher positions on the leaderboards. In scenarios where theoretical understanding lags behind empirical results, a new state-of-the-art (SOTA) usually emerges from trial and error experimentations often producing quite arbitrary heuristics. Faced with the theoretical lacuna, practitioners confront the temptation of post-hoc speculations that might assume the role usually played by theoretical explanations (cf. Lipton and Steinhardt, 2019). When occurring alongside (accidental) misattributions of the sources of empirical gains, f.e. reporting improvements from neural architecture changes when, in reality, they stem from hyperparameter tuning (ibid.), the following might ensue. Instead of achieving the epistemic aim of understanding generalisations, leaderboards might merely encourage post-hoc hypotheses fitted to the results of quite arbitrary heuristics. The incredibly fast pace of the leaderboards climbs, natural language processing (NLP) is among the best of present examples (cf. Strubell et al., 2019 for an estimation of the SOTA NLP's environmental costs), makes such bad practices a siren song, which could considerably hamper the discipline's twofold epistemic goal.

2. The Future Shape of Human-Artificial Intelligence Nexus and its Environmental Costs

ML of generalisations pursues minimisation of empirical risk that should guarantee accurate inferences regarding the task at hand. Generalisation learning thus seeks to minimise the inductive risk associated with the task. From the body of theoretical approaches to induction, Norton's material theory (2003), positing that inductive inferences are grounded in local facts holding in particular domains (ibid.), shows a potential to illuminate the future shape of Human-Artificial Intelligence Nexus. It is plausible to argue that a successful ML of a generalisation produces an inductive schema and minimises its risk by tethering it to the local facts found in the training data. As per Norton, successful ML might be thus epistemologically explained and justified by favouring a myriad of local inductive schemas over a few elusive global (universal) ones.

Regardless of whether concerning *human* or *machine* learning of generalisations, any minimisation of empirical risk by way of localising the inductive schema involves environmental costs. It can be argued that ML experiments achieving generalisations create local inductive schemas underwritten by local facts (from the evidence, i.e. training data), which enable the minimisation of empirical risk. Similarly, for humans, as argued by Norton (2003), there are no universal inductive schemas and inductive inferences hold only in particular domains, being underwritten by local facts. A wider and deeper apprehension of local facts extends the cognitive reach of humans as well as of machines. The crucial distinction is that humans, due to the evolutionary pressures, remain frugal learners compared to the sample inefficient ML (e.g. cf. the recent OpenAI Five [OpenAI, 2019] experiment that required 45,000 human years of training to defeat the best human players of the Dota 2 computer game).

Arguably, it's not an exaggeration to posit that humanity's epistemic endeavour is going through a period of unprecedented transformations. In a while, it might so happen that out of the quantifiable total of inductive inferences most will be carried out by 'tools-for-optimal-action'. More importantly, the total number of inferences will most likely skyrocket, as humanity will eagerly boost the languishing bits of its scientific and common epistemic endeavours alike, and on top of them quite probably invent new ones. As a result, the environmental costs of the anthropic and artificial localisations of inductive risk, which correspond to the total of inductive generalisations, will soar as well. Compared to the human inductive inferences, inductive schemas furnished by ML generalisations will, however, come at a considerably higher price if the practice remains harmful to the environment.

3. The Environmental Costs of Machine Learning Research

From a bird's-eye view, utilising accuracy to measure SOTA, while incentivising leaderboard climbs by any means available, constitutes ML of generalisations as an environmentally hurtful endeavour. An improved accuracy can be purchased by additional computational resources, which enable a threefold growth that typically leads to a new SOTA accuracy (Schwartz et al., 2019). Computational resources, however, come at an environmental cost, which becomes significant as the relationship between the experiments' scale and the gained accuracy turns less favourable (an exponential growth of the experiment for an approximately linear gain in accuracy, cf. ibid.). Schwartz et al. (2019) formalised the cost of climbing the accuracy leaderboards as the experiments' threefold growth:

$$\mathtt{Cost(Result)} \propto \mathtt{E \cdot D \cdot H} \quad \text{(Equation 1)}$$

The total cost of a ML experiment (result) grows linearly with increasing (E), the cost of processing a single example, (D), the volume of training data, and (H), the number of the experiment's variants executed to find a new SOTA accuracy (ibid.). The unfavourable relation between cost and accuracy captured by Eq. 1 is empirically discernible as the macro-trend of growing computational resources consumption in ML (Amodei and Hernandez, 2018; Sastry et al., 2019) as well as the growing complexity of ML experiments. The complexity increases with the number of times an initial ML model is retrained to find a good fit of the model's architecture and hyperparameters to data, i.e. to the task at hand, thus reaching a new SOTA accuracy. In setups where such a search over architecture/hyperparameter spaces is guided by human researchers, we are observing thousands of training cycles per experiment (the third term of Eq. 1; cf. Strubell et al., 2019). In setups where the human guidance is replaced with artificial evolutionary search over architecture/hyperparameter spaces (i.e. Neural

Architecture Search [NAS]), the number of trained models per experiment reaches tens of thousands (cf. So et al., 2019; Real et al., 2019). The rest of this section, utilising Strubell et al.'s trailblazing analysis (2019), seeks to dispel a well-intentioned yet ultimately incorrect notion which assumes that a single training run of an already developed ML model on a benchmark dataset can be in some way representative of the environmental costs associated with ML research.

To better illustrate the point, we might look at one of the most prominent leaderboards, ImageNet – an image classification challenge, and a landmark ML model. Even though it is now possible to train ResNet-50 (He et al., 2016) to the accuracy once ranking high on the ImageNet leaderboard in 2:43 minutes optimising for speed or for $12.60 optimising for cost (DAWNBench, 2019), this does not mean that better hardware and/or training techniques attenuated the costs of leaderboard climbs. Rather, reaching the higher rungs, and thus accuracy, requires larger models and more data. In case of fixed benchmark datasets such as ImageNet, which are required for leaderboards, this means acquiring data for 'pretraining', allowing to gain, combined with other improvements, higher accuracy. Touvron et al. (2019) used 940 million public images for a weakly-supervised pretraining experiment on ResNeXt-101 32x48d architecture, comprising 829 million parameters (versus ResNet-50's 25.6 million parameters), and fine-tuned the result to ImageNet. The experiment now (January 2020) ranks fourth on the ImageNet leaderboard (Papers With Code, 2019a). The resources required for performing such an experiment are incomparable to a single training run of an already developed, standard ML model (e.g. ResNet-50) on a benchmark dataset. The pursuit of the top places on leaderboards clearly translates into larger experiments (ML models and data alike), growing computational demands, and thus increasing of environmental impacts.

Therefore, even if we observe progress in the base-level efficiency, it is not an indicator that ML's environmental impacts are negligible.

Expressed in terms of the macro-trend, Amodei and Hernandez (2018) showed that since the beginning of the Deep Learning (DL) epoch in ML, the amount of computational resources spent on the largest experiments, which correspond to the top places in leaderboards, doubles every 3.4 month. Since 2012, when AlexNet (Krizhevsky et al., 2012) first dominated non-DL methods on the ImageNet leaderboard, the amount of compute has grown by more than 300,000x, and is presently reported as days at which the experiment ran at a petaflop/s (floating point operations per second, Amodei and Hernandez, 2018) on multiple GPUs or TPUs (Graphics Processing Unit, Tensor Processing Unit, the spent petaflop/s-days grow more rapidly with an increasing parallelisation of the experiments). The scale ranges from petaflop/s-days of running time to lower or even upper hundreds (this translates into tens of thousands of years in the human temporal frame of reference, OpenAI, 2019), in case of the most demanding experiments reaching over a thousand of petaflop/s-days (Amodei and Hernandez, 2018, this is, however, the present-day upper-bound, which is not representative of a typical experiment). Amodei and Hernandez (2018) suggested that at least in short-term the macro-trend is likely to continue since we have not exhausted the room for improvement in the flop/s per Watt ratio as well as the opportunities for a better utilisation of parallelism in ML experiments. Therefore, considering only economic constraints, achieving a greater base-level efficiency, i.e. a quicker/cheaper single training run of an already developed ML model on a benchmark dataset, is likely to lead to growing experiments, pursuing higher accuracy and escalating the leaderboard climbs. As we find ourselves still close to the beginning of the trend, it is timely to estimate CO_2 emissions stemming from ML experiments. The emissions derive from the energy consumed to satisfy computational demands of the experiments and can thus approximate ML's contributions to the anthropogenic change of the Earth system.

Strubell et al. (2019) estimates the power consumption of a ML experiment as a sum of power used by the hardware multiplied by additional power requirements for sustaining the infrastructure (i.e. cooling etc.). The amount of CO_2 released by the experiment is calculated by multiplying the sum of power usage by the average CO_2 emission per kilowatt-hour provided by the U.S. Environmental Protection Agency (ibid., EPA). Strubell et al. (2019) then compares selected types of ML experiments to the average CO_2 released during a round-trip flight from New York to San Francisco (1 passenger, *1984 lbs* [*900 kg*]), 1 year of average human life/1 year of average American life (*11,023 lbs* [*5 tons*] / *36,156 lbs* [*16.4 tons*]), or an average car's lifetime including the consumed fuel (*126,000 lbs* [*57.2 tons*]). Taking ML in Natural Language Processing (NLP) as a case study (ibid.), the magnitude of the CO_2 emission derives from whether the search over the architecture and hyperparameter spaces were guided by human researchers or by an artificial evolutionary search. Considering an example of the former kind of experiment, provided by Strubell et al.'s (2019) own account based on developing a novel NLP model, the amount of CO_2 emissions is estimated at *78,468 lbs* (*35.6 tons*, calculated using the U.S. average of CO_2 emissions per kWh published by EPA). The emissions accumulated from 4,789 ML models trained during the experiment (the third term of Eq. 1), which led to the best attuned model constituting the SOTA on some of the Semantic Role Labelling leaderboards (Papers With Code, 2019b). For an NLP example of the latter kind of experiment, based on neural architecture search, i.e. evolutionary search for the best model, Strubell et al. (2019) estimates the amount of CO_2 emissions at *626,155 lbs* (*284 tons*, using the same estimation method as above). Perhaps as a genuine cautionary tale can serve a recent experiment (Meng et al., 2019) claiming to utilise 512 GPUs for three straight months to pretrain a machine translation model on roughly 40 billion sentence pairs. The pretraining brought only a modest SOTA improvement on the benchmark dataset, the experiment lacking a foundational contribution overall, and the large set used for pretraining was not released yet. As the experiment cannot be repeated to verify the results, and it is an open question whether it should be repeated at all, the only surviving, tangible result is a likely substantial CO_2 release. Finally, it needs to be emphasised that the results of ML experiments sometimes lack transferability. This means that retraining/fine-tuning is required to deploy the ML models in new domains or to accomplish new tasks. Therefore, further CO_2 emissions are to be expected (cf. Strubell et al., 2019).

Compared to CO_2 emissions associated with common areas of human life, the amount of CO_2 released by ML experiments is non-trivial. More so, considering the recent explosive development of ML as a discipline/field, which can be illustrated by the growing number of research papers. At the end of 2018, the number of submissions to ML-related sections of arXiv.org (a popular repository of e-preprints [not peer reviewed] operated by Cornell University) reached 3,000 papers per months (Dean, 2019). A similarly explosive trend in the number of submissions accompanies academic ML conferences (the leading academic ML conference NeurIPS quadrupled over the last five years, in 2019 reaching 6,743 submissions, Beygelzimer et al., 2019), albeit not all the experiments are as demanding as the above examples.

Facing the reality of behemoth experiments, which have become a standard practice for improving the accuracy, the discipline realised that its epistemic aim, the quest for theoretical and empirical understanding of generalisation, comes at environmental costs. The ML community offered two possible remedies. First, the computational resources required for climbing the accuracy leaderboards should be powered only by energy from renewable sources, thus securing 100% sustainability (cf. Hölzle, 2019). However, the claims of consuming 100% renewable energy usually refer to the consumption per annum, but the real consumption at certain times, e.g. at night-time for solar, is still satisfied by burning of fossil fuels (de Chalendar and Benson, 2019). The true 100% renewable consumption would require storing the energy surplus generated during the peaks of renewable energy supply. Second, publication venues should explicitly reward the research that reduces any of the Eq. 1 quantities while securing a competitive, *albeit non-SOTA*, accuracy (cf. Schwartz et al., 2019, a related work also suggests reporting the validation results obtained during training to allow the estimation of a computational budget required for a given validation accuracy, cf. Dodge et al., 2019).

4. The Environmental Costs of Gratuitous Generalisation Capabilities

A foresight of potentially sobering environmental effects arising from the discipline's epistemic aim, and the willingness to confront them, is indeed laudable. Yet the picture of the environmental impacts linked to ML remains incomplete. It's merely the tip of the iceberg, the rest corresponds to the paradigm of generalisation learning which optimises the 'tools-for-optimal-action' exclusively for accuracy. To obtain a faithful picture of the environmental impact, apart from the cost of research indicated by Eq. 1, it would be necessary to factor in *also* the cost that accumulates while using the tool's generalisation capability to accomplish the designated task. First, to estimate the energy consumed by a single application of the tool's generalisation capability, we would need to establish the number of computational operations per inference. Second, this quantity should be multiplied by the number of inferences which are expected to be performed by all future instances of that particular 'tool-for-optimal-action' and appended as the fourth member to Eq. 1.

$$\mathrm{Cost(Result)} \propto E \cdot D \cdot H \cdot cI \quad \text{(amended Equation 1)}$$

In the amended Eq. 1, c stands for the number of computational operations per inference and I represents the total number of future inferences (as per above). Only after such an amendment would Eq. 1 begin to converge on the tool's true environmental impact. In this context, it needs to be emphasised that with increasing accuracy grows also the time required for performing an inference (cf. Bianco et al., 2019), thus increasing the amount of computational resources, and energy, required for deployment. Although some encouraging results emerged recently, showing a progress in the accuracy to inference latency ratio (cf. Gupta and Tan, 2019), the second and third term of Eq. 1, which are essential for the progress, remain expensive.

It's safe to assume that our eagerness to let ML take care of even the most negligible everyday tasks will eventually lead to the *all-pervasive use* of generalisations provided by the 'tools-for-optimal-action'. Such prospect, however, hampers even a ballpark estimate of the fourth quantity, which unfortunately renders the amended version of Eq. 1 impractical. Reaching an impasse, the issue clearly requires a different kind of approach that might emerge from the following shift of perspective.

Apart from minimising the value of Eq. 1, a principled approach would entail a redefinition of the 'tool-for-optimal-action' concept itself. Rather than optimising a single objective, i.e. accuracy, the 'tool-for-optimal-action' would be required to observe a task-specific limit stipulating the maximum number of computations per inference. Such constraint would throttle the tool's energy consumption and attenuate its environmental footprint. It would also diminish the accuracy of its generalisation capability. In this regard, the second objective aims to reflect the fact that even a diminished level of generalisation capability might furnish an optimal tool. For example, the task at hand might include a human in the loop providing corrections (typical for cognitive extending), or a mistake remains so cheap that a more powerful tool could not be justified.

Expressed formally, holding all cognitive tasks equally demanding of generalisation, while optimising the 'tools-for-optimal-action' solely for the SOTA accuracy, creates an aggregate surplus of the generalisation capability. The underlying single-objective generalisation learning arises from an uneven epistemic aim, which rewards climbing the accuracy leaderboards. This twofold interplay then produces *gratuitous* generalisation capabilities, which threaten to become dissipative and thus pernicious to the environment. It could be argued that the multi-objective generalisation learning could at least partially alleviate such adverse effects. By way of assessing the cognitive complexity of any task at hand, it should be possible to base the 'tools-for-optimal-action' on a favourable trade-off between the energy consumption and generalisation capability, while also encouraging reductions of the research cost indicated by Eq. 1. Put differently, by agreeing on a reasonable accuracy, which approximates the cognitive demands of a particular task, it becomes feasible to limit the computational budget of the generalisation learning (Eq. 1) and of the deployment-time inferences accordingly.

Anticipating the likely near-term all-pervasive use of ML generalisations, the multi-objective definition guarantees that the 'tools-for-optimal-action' achieve the stipulated accuracy without possessing gratuitous generalisation capabilities at the environment's expense. Two shifts in the discipline's epistemic aim could contribute to such outcome. First, the race to the top of the leaderboards is perhaps better abandoned, unless it ceases to depend on infinitesimal improvements purchased by inflating the quantities of Eq. 1. Second, the discipline might be near the verge of discovering theoretical foundations which would turn the future generalisation learning into a principled effort (Nagarajan and Kolter, 2019; Frankle and Carbin, 2019; Jiang et al., 2019). In that case, it would be most likely possible to satisfy the cognitive as well as environmental objective while also significantly reducing the research cost indicated by Eq. 1.

5. Conclusion

A full realisation of such scenario would see the techno-sphere (Haff, 2014), a part of the Earth System that sustains modern civilisation and its inhabitants, grow an epistemic dimension. Anticipated by the present crave for ML/AI cognitive extending, the future of humanity's scientific and everyday endeavours would then likely become predicated upon an enormous amount of non-anthropic inductive inferences. If this future Human-Artificial Intelligence Nexus remains dependent on environmentally harmful (ineffective) ML/AI, the techno-sphere's epistemic dimension will likely exacerbate the perils of Anthropocene, which deteriorate the Earth System. By fleshing out this possible shape of Human-Artificial Intelligence Nexus to come, we hope to show that perhaps the greatest rejuvenation of our epistemic endeavour since the scientific method might come at environmental costs, if the initiatives like GreenAI fall on deaf ears.

The paper proposed a techno-philosophical way of thinking about future environmental costs of ML/AI, which seeks to offer an alternative to often provocative normative opinions. It is argued that if ML/AI research and applications remain ineffective, we should be prepared for unforeseen environmental costs. The ineffectiveness lies in the single-objective learning, pursuing the best possible accuracy at all costs. A possible solution could be based on a multi-objective learning, where the first objective, a *reasonably defined* task at hand, provides the upper-bound for the second objective, accuracy.

References

Amodei, D., Hernandez, D. (2018). AI and Compute. https://openai.com/blog/ai-and-compute/.

Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D. (2019). Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893v2*.

Bartlett, P. L., Foster, D. J., Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Beygelzimer, A., Fox, E., d'Alché-Buc, F., Larochelle, H. (2019). What we learned from NeurIPS 2019 data. https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2019-data-111ab996462c.

Bianco, S., Cadene, R., Celona, L., Napoletano, P. (2018). Benchmark Analysis of Representative Deep Neural Network Architectures. *IEEE Access* 6, 2169-3536.

DAWNBench (2019). DAWNBench: An End-to-End Deep Learning Benchmark and Competition. https://dawn.cs.stanford.edu/benchmark/ImageNet/train.html.

Dean, J. (2019). Machine Learning Arxiv Papers per Year. https://twitter.com/JeffDean/status/1135114657344237568.

de Chalendar, J. A., Benson, S. M. (2019). Why 100% Renewable Energy Is Not Enough. *Joule* 3(6), 1389-1393.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., Smith, N. A. (2019). Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,* 2185-2194, arXiv preprint arXiv:1909.03004.

Frankle, J., Carbin, M. (2019). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *7th International Conference on Learning Representations*.

Gupta, S., Tan, M. (2019). EfficientNet-EdgeTPU: Creating Accelerator-Optimized Neural Networks with AutoML. https://ai.googleblog.com/2019/08/efficientnet-edgetpu-creating.html.

Haff, P. (2014). Humans and technology in the Anthropocene: Six rules. *The Anthropocene Review* 1(2), 126-136.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*.

Hölzle, U. (2019). 100% renewable is just the beginning. https://sustainability.google/projects/announcement-100/

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., Bengio S. (2019). Fantastic Generalization Measures and Where to Find Them. In *8th International Conference on Learning Representations*.

Kawaguchi, K., Kaelbling, L. P., Bengio, Y. (2019). Generalization in Deep Learning. *arXiv preprint arXiv:1710.05468*.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*.

Lipton, Z., Steinhardt, J. (2019). Troubling Trends in Machine Learning Scholarship. *ACM Queue* 17(1).

Meng, Y., Ren, X., Sun, Z., Li, X., Yuan, A., Wu, F., Li, J. (2019). Large-scale Pretraining for Neural Machine Translation with Tens of Billions of Sentence Pairs. *arXiv preprint arXiv:1909.11861v3.*

Nagarajan, V., Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Neyshabur, B., Bhojanapalli, S., McAllester, D., Srebro, N. (2017). Exploring Generalization in Deep Learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Norton, J. D. (2003). A Material Theory of Induction. *Philosophy of Science* 70(4), 647-670.

OpenAI (2019). OpenAI Five Defeats Dota 2 World Champions. https://openai.com/blog/openai-five-defeats-dota-2-world-champions/.

Papers With Code (2019a). Image Classification on ImageNet. https://paperswithcode.com/sota/image-classification-on-imagenet.

Papers With Code (2019b). Linguistically-Informed Self-Attention for Semantic Role Labeling. https://paperswithcode.com/paper/linguistically-informed-self-attention-for.

Real, E., Aggarwal, A., Huang, Y., Le, Q. V. (2019). Regularized Evolution for Image Classifier Architecture Search. In *33rd AAAI Conference on Artificial Intelligence*.

Sastry, G., Clark, J., Brockman, G., Sutskever, I. (2019). Addendum to AI and Compute. https://openai.com/blog/ai-and-compute/#addendum.

Schwartz, R., Dodge, J., Smith, N. A., Etzioni, O. (2019). Green AI. *arXiv preprint arXiv:1907.10597*.

So, D. R., Liang, C., Le, Q. V. (2019). The Evolved Transformer. In *Proceedings of the 36th International Conference on Machine Learning*.

Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In *57th Annual Meeting of the Association for Computational Linguistics*, arXiv preprint arXiv:1906.02243.

Touvron, H., Vedaldi, A., Douze, M., Jégou, H. (2019). Fixing the train-test resolution discrepancy. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. (2017). Understanding Deep Learning Requires Rethinking Generalization. In *5th International Conference on Learning Representations*.