Aligned deep learning as a random sampling method

Leonardo Pedro

December 25, 2023

Abstract

We show that the alignment problem in deep learning can be separated in two: A) estimate systematic uncertainties as in Engineering; B) new due to Big Data, how to randomly sample computable models from a set of Bayesian models, almost all uncomputable. Deep neural nets solve B). Since in most sufficiently complex systems, it is unlikely to find a sufficiently simple statistical metric that it is not misleading (or rogue, for instance the GDP in economy), the (dis)alignment is not a problem but a feature: if the estimate A) would not be likely misleading, then the sampling B) would not be random/unpredictable and if it would not be random it would not generalize well (by definition of generalization in a Bayesian context).

1. Introduction

Heuristics are simple strategies to quickly find solutions to complex problems. Since both heuristics and mathematics are used in problem-solving, often the same word (in the context of problem-solving) has a mathematical definition and a heuristic meaning which are different.

Deep learning[1] and Bayesian inference represent different classes of heuristics to solve inverse problems. But, there are no detailed studies about the relation (or just the differences) between the mathematical definitions of deep learning and Bayesian inference[2]. This is surprising, since deep learning has several properties (for instance, "good generalization") which have a mathematical definition in the context of Bayesian inference. Sure, the heuristics usually associated with Bayesian inference (for instance, Bayesian neural networks) are different from Deep neural networks[1], but that does not imply much from a mathematical point of view.

In fact, there is experimental evidence that perception in the human brain is consistent with Bayesian inference[3]. Moreover, trial-and-error with different Deep learning methods and experimental/empirical results, uses Bayesian inference (in the mathematical sense)[4]. Note that such trial-and-error is crucial in most applications of Deep learning, thus it cannot be separated from Deep learning itself:

> *"However, actually achieving good performance from a deep learning model still requires some finesse and application of various heuristics. It is safe to say that a significant amount of the practice of deep learning remains more art than science. The good news is that once effective heuristics for a particular problem domain have been developed, these same heuristics can often be applied with little modification to other problems in the same domain."*[5]

The conventional wisdom is that in the past, the research in neural networks was neglected due to theoretical prejudice[6] and therefore, theory should now somehow be neglected to free resources for "Engineering science: inventing new artifacts" (whatever that means). However, whatever methods we use for "inventing new artifacts" they always involve trial-and-error which can always be considered an approximation to Bayesian inference. In particular, there are no non-informative priors in Bayesian inference[7], therefore theoretical prejudice is unavoidable (even in "Engineering science"). The way out is to be aware that there is no trial-and-error without theoretical prejudice, thus we should not discard patterns discovered by other people just because they do not fit our prior assumptions (as it happened in the past when society neglected neural networks, despite there were already patterns of consistent good results).

Other than assumptions, there is no good reason to believe that the consistent good results of neural networks are unrelated with Bayesian inference. In fact, deep neural networks implement extremely non-linear functions which makes them unpredictable (just like pseudo-random generators) which leads to pseudo-randomness and Bayesian inference. In this article we will show that the alignment problem[8][9] in deep learning can be separated in two: A) estimate systematic uncertainties as in Engineering; B) new due to Big Data, how to randomly sample computable models from a set of Bayesian models, almost all uncomputable. We also present evidence that Deep neural nets solve B).

Assuming that everything relevant about Artificial Intelligence is emergent and new/empiric is convenient, since it deflects responsibility[10] and it allows private appropriation of public goods/knowledge[1]. These assumptions are more easily validated by the Media and Software Industries, since they are convenient for those who already have a dominant position in these Industries[2]. For instance, suppose a doctor applies the wrong dosage of a medicine to treat a mental disorder in a

[1]For instance, https://en.wikipedia.org/wiki/Dilution_(neural_networks)#cite_ref-9 : "*The patent is most likely not valid due to previous art. "Dropout" has been described as "dilution" in previous publications. It is described by Hertz, Krogh, and Palmer in Introduction to the Theory of Neural Computation (1991), pp. 45, Weak Dilution. The text references Sompolinsky The Theory of Neural Networks: The Hebb Rules and Beyond in Heidelberg Colloquium on Glossy Dynamics (1987) and Canning and Gardner Partially Connected Models of Neural Networks in Journal of Physics (1988). It goes on to describe strong dilution. This predates Hinton's paper.*"

[2]See for instance John Mearsheimer (2023) in https://youtu.be/t2451jFeZp0?t=2497 from 41:37 until 44:18.

patient, would he get away with "these brain effects are emergent"? Couldn't a Pharmaceutical Lab earn much more money if it would be allowed to skip the Clinical Trials of a revolutionary medicine for the brain, with "the Clinical Trials use slow Bayesian methods, we have our own new/empirical methods which are much faster"?

The public information space is messy for the same reasons that the scientific information publicly available is messy. Peer-review, the scientific method, intellectual abilities (including artificial intelligence) and authority, shared values, rules and incentives within a large community cannot solve such problem, otherwise they could be used to solve the similar problem of the public information space. Half-truths (and even explicit lies) are abundant in the scientific information publicly available. Most scientists do not communicate clearly on purpose, not because they want to deceive society, but simply to avoid serious or unnecessary problems for their careers and families. Just like any other person who has to speak in public (a journalist, for instance) about any other subject.

2. Systematic uncertainties and Bayesian priors

> *"The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted – recognizing a face, lifting a pencil, walking across a room, answering a question – in fact solve some of the hardest engineering problems ever conceived... As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come."*[11]

Moravec's paradox is the observation by artificial intelligence and robotics researchers that, contrary to traditional assumptions, reasoning requires very little computation, but sensorimotor skills require enormous computational resources (due to the need of enormous amount of previous knowledge). The most difficult human skills to reverse engineer are those that are *unconscious*.

> "*The deliberate process we call reasoning is, I believe, the thinnest veneer of human thought, effective only because it is supported by this much older and much more powerful, though usually unconscious, sensorimotor knowledge. We are all prodigious olympians in perceptual and motor areas, so good that we make the difficult look easy. Abstract thought, though, is a new trick, perhaps less than 100 thousand years old. We have not yet mastered it. It is not all that intrinsically difficult; it just seems so when we do it.*" [12]

Reasoning is the capacity of consciously applying logic to seek truth and draw conclusions from new

or existing information. Reasoning can be automated, and thus be aided by a machine external to the biological brain. Automated reasoning is considered a sub-field of artificial intelligence, tools and techniques of automated reasoning include the classical logics and calculi, fuzzy logic, Bayesian inference, reasoning with maximal entropy and many less formal ad hoc techniques.

If we define Machine learning as "*the study of computer algorithms that can improve automatically through experience and by the use of data*" (as in Wikipedia), then machine learning is a subcase of automated reasoning applied to computer algorithms, that is, the capacity to automatically produce new computer algorithms from new and existing information. For instance, since the supervised learning problems in machine learning can be thought of as learning a function from examples, these problems can be formulated as Bayesian modeling[13].

For a class of hard problems (such as some sensorimotor skills), Machine Learning automatically produces a *probably* correct computer algorithm (through automated reasoning) instead of leaving to the human programmer the daunting task of finding himself a correct algorithm for the hard problem.

In the presence of a finite (eventually arbitrarily large) number of random variables[3], Bayesian inference is by far the most general and theoretically grounded framework available to formulate automated reasoning[14][15]. The systematic uncertainties in Engineering can be defined as Bayesian prior probability distributions[16], which describe previous knowledge about a system.

Then a brain (natural or artificial) has two parts: 1) a symbol grounding system[17](unique for each brain) which transforms signals coming from the real-world into inputs for the second part of the brain, and transforms the output from the second part of the brain into real-world actions; 2) a system (which apart from differences in performance and previous knowledge, it does the same in all brains) doing Bayesian learning in a standard probability space from the abstract inputs coming from the second part of the brain, where the posterior probability is the output. Bayesian inference requires systematic uncertainties/previous knowledge, which are often hard to estimate.

The problem A) estimate systematic uncertainties as in Engineering, is often addressed by dividing tasks between specialized modules[18], such that the modules with leadership/coordination tasks (or other tasks where it is hard to estimate systematic uncertainties) have no effective power, while the modules with effective power have tasks where good estimates for the systematic uncertainties exist. This happens also in most human societies, where each module is a person. Note that problem A) is often not completely solved, and consequently safety requires a significant cost on the performance of the system. Moreover, problem A) is task-specific, that is, the systematic uncertainties are different

[3]Note that in the presence of an infinite number of random variables, Bayesian inference struggles also theoretically, we study such case in another article.

for each task and may require different methods to estimate them, but this does not imply that Bayesian inference somehow is not good enough.

As we will see when discussing problem B), humans and deep neural networks do not solve problems in fundamentally different ways, although quantitatively things may be different (a computer is often faster than a human in arithmetic, for instance). Therefore, the problem A) for AI is not fundamentally different from problem A) for Engineering or Social Management. Consider for instance, how to manage the surveillance and eventual counter-measures about nuclear attacks: no single person can be entrusted with such amount of responsibility and power, no matter his apparent alignment with human values. However, sharing the responsibility over a too large group of persons potentially hurts the speed of the decisions, this may be critical. Thus, there is a trade-off between ensuring that our people do not commit irresponsible actions, and empowering them enough to defend us effectively. The same kind of trade-off exists if we add artificial intelligence to the mix, except now the speed of the decisions can be much faster.

Note that once we estimate the prior probability/systematic uncertainty, its logarithm can be added to the function to be maximized in deep learning, thus the main trouble is how to estimate such prior probability. Also, Bayesian inference can be defined as the decision process under uncertainty which minimizes losses when betting against an adversary[7], so we could also say that Bayesian inference mimics generative adversarial networks[19].

3. Bayesian inference in the presence of Big Data

From a mathematical point of view, Bayesian inference is still valid in the presence of Big Data, since it is valid for an arbitrarily large number of random variables, as long as it is a finite number.

However, Bayesian inference is often not tractable due to the normalization term in the Bayes' theorem[15], therefore approximation inference techniques must be used often. Neural networks and other models in machine learning can be considered as ad hoc approximations to Bayesian interference, which often have better performance than other more theoretically grounded approximations.

The so-called *catastrophic forgetting* characteristic of neural networks[20] is consequence of the fact that neural networks only produce deterministic predictions: the predictions for which there is a good degree of belief and therefore should not be erased are not distinguished from those for which there is not a good degree of belief and thus can be modified. Therefore, neural networks follow the Bayesian updating rule (which is derived from basic logic principles) only approximately and unpredictably.

Until 2008[21], there were no alternative approximations of Bayesian inference which were competitive

with neural networks in big data, that is, for deterministic functions with much more than a few thousands of parameters. But today there is not such superiority of deep neural networks with respect to all other approximations to Bayesian inference, even for deterministic functions with billions of parameters[22][23][24][25][26][27]. Ironically, we can even use deep neural networks to do meta-learning and find approximations to Bayesian inference which are better than deep neural networks themselves[1].

Deep neural networks (or improvements of them[28]) are not a "model that mysteriously generalizes well", instead they are a sampling method that solves problem B), how to sample the space of solutions with high enough likelihood (near but not exactly the maximum likelihood[29], this is often the space that a perfect sampling method would effectively cover, due to the Wilks theorem) conditioned on having low computational complexity. The low computational complexity is essential in Big Data, these are the solutions that are effectively computable. Often, the computable solutions spread approximately homogeneously over the space of all solutions, so that sampling only computable solutions produces similar results to sampling all solutions. Moreover, often (not always) the prior assumption that the solution is effectively computable makes sense, for instance when we are trying to mimic human skills (these are often effectively computable) or when we are searching for "patterns" in data, that is, "simple" explanations for what is happening.

Note that the optimization problem improves for deeper neural networks[30], that is, for a deep enough neural network often the training converges to a local maximum that is near the maximum likelihood (but not exactly). Moreover, many improvements[28] (and even alternatives[31]) to neural networks may also solve problem B), the key requirement is that the sampled solutions spread approximately homogeneously over the space of all solutions with an acceptable likelihood.

One interesting application of the framework described above, is to the problem of understanding why deep neural networks often generalize well: we accept all explanations as long as it is simple enough, it is consistent with the known evidence, and it contributes to correct predictions about future evidence. This is also the effective prior distribution covered by deep neural networks.

Neural networks can then be distilled[32] (for instance, to Wavelets, or a Krylov subspace, or Explainable AI[33]) and thus, they are part of the sampling algorithm, they do not necessarily define the model, that is, the space of all possible solutions, they only define which possible solutions have low computational complexity.

Conjecture/thought experiment/trial/speculation/hypothesis/ansatz/religion (in the sense that it fills the unknown) can be integrated with prior knowledge, using a parametrization of the prior distribution as input, and using a sampler (with a random input or latent variable) which not only will extract a number out of the prior knowledge, but it also contributes to the robustness of several

conjectures. That is, if many conjectures (combining random initialization and also a random input) about the same question produce consistent answers, then the model "knows" the answer, otherwise it doesn't.

Many learning machines can certainly be redefined as a classical Hamiltonian system[34], and thus be related to physics. More generally, the optimization is a sequence of actions on a system, this can be redefined as a Quantum Hamiltonian acting on a wave-function which parametrizes an ensemble of systems. Not only can deep neural networks be parametrized by a Quantum system, but also the other way around: any quantum system can be parametrized by a deep neural network, up to an arbitrarily small error, in the following way. From reference[5] and the universal approximation theorem for probability distributions[35][36][16], we know that a deep neural network can be used to guess/conjecture/hypothesize the full definition of a high-dimensional (eventually infinite-dimensional) wave-function, from a large but manageable number of parameters. But hypothesis are part of probability theory (think about statistical sampling or even the scientific method itself) and probabilities of probabilities are crucial in (classical and quantum) field theory. Thus machine learning is related to physics through probability theory which is related to (classical and quantum) physics through the quantum formalism, all in a fundamental way with many real-world applications.

While it is true that neural networks solve some important problems that symbolic AI cannot solve, they are not that different from symbolic AI since several popular types of neural networks are Turing complete[37] (including a version of Transformers that uses rational numbers with arbitrary precision). Thus, a sufficiently complex neural network can be defined as a symbolic AI which allows differentiable programming[38], so it is fast to optimize.

4. Why deep neural networks do not overfit

Machine Learning (for instance Deep Neural Networks) is not firmly based in probability theory. In Machine Learning, methods inspired by probability theory are used often[39], but the formalism is based in approximations to deterministic functions, guided by a distance (or equivalently, an optimization problem) and not a measure. In fact, two of the main open problems are the alignment of models and the incorporation of prior knowledge[1], which could be both well solved by a prior measure if there would be any measure defined.

Under reasonable assumptions, almost all functions are not computable not even approximately. Thus, Machine Learning works because the functions we are approximating are in fact probability distributions (eventually after some reparametrization[40]). This shouldn't be surprising, since Classical Information Theory shows (under reasonable assumptions) that probability is unavoidable

when we are dealing with representations of knowledge/information[41][39]. But in Machine Learning the probability measure is not consistently defined (despite that many methods are inspired by probability theory), the probability measure emerges from the approximation[40] and often in an inconsistent way. The inconsistency is not due to a lack of computational power since modern neural networks can fit very complex deterministic functions and fail badly[42][43] in relatively simple probability distributions (e.g. catastrophic forgetting or the need of calibration to have some probabilistic guarantees[43]).

This unavoidable emergence of a probability measure should be investigated as a potential source of inefficiency, inconsistency and even danger. If the emergence of a probability measure is unavoidable, why don't we just define a probability measure in the formalism consistently? Many people say "it is how our brain works", so mathematics should step aside when there is empirical evidence.

But the empirical evidence is: oversized deep neural networks still generalize well, apparently because often the learning process converges to a local maximum (of the optimization problem) near the point where the learning begun[44]. This implies that if we repeat the learning process with a random initialization (as we do when we consider ensembles of neural networks[42][45]), then we do not expect the new parameters to be near any particular value, regardless of the result of the first learning process. This expectation is justified by the fact that every three layers of a wide enough neural network is a universal approximation of a function[46], so any deviation introduced by three layers can be fully corrected in the next three layers, when composing dozens or hundreds of layers as we do in a deep neural network. Then the correlation between the parameters corresponding to different local maximums converges to zero, when the number of layers increases, because extremely non-linear functions are unpredictable (just like pseudo-random generators) which leads to pseudo-randomness.

Thus, there is empirical evidence that oversized deep neural networks still generalize well, precisely because a prior measure emerges: deep learning does not converge to the global maximum and instead to one of the local maximums chosen randomly, effectively sampling from a prior measure in the sample space defined by all local maximums. This is consistent with the good results achieved by ensembles of neural networks[42][45], which mimic many samples. However, it is a prior measure which we cannot easily modify or even understand, because the measure space is the set of all local maximums of the optimization problem. But, since we expect the parameters to be fully uncorrelated between different local maximums, then many other prior measures (which we can modify and understand, such as the uniform measure) should achieve the same level of generalization.

This is not a surprise, since oversized statistical models that still generalize well were already found many decades ago by many people[47]: a standard probability space with a uniform probability measure can be infinite-dimensional (the infinite-dimensional sphere[47], for instance).

Note that generalization is often defined independently of the problem of overfitting, but these are both a subproblem of whether or not the posterior measure is compatible with some prior measure. Despite that generalization is often defined with respect to "future" data, we know nothing about future data except if we make assumptions which then define a prior measure, for instance by assuming that future data will be similar to data we already collected but did not use to estimate the posterior measure, such unused data defines a prior measure. Overfitting is when we choose a set of models which have a likelihood which is too close to the maximum possible, while our prior includes likelihoods close but not too close from the maximum, thus the posterior measure (that is, the set of models) is unlikely ("*too good to be true*") with respect to the prior measure.

More empirical evidence: no one looks to a blurred photo of a gorilla and says with certainty that it is not a man in a gorilla suit. We all have many doubts, when we are not sure about a subject we usually express doubts through the absence of an action (not just us, but also many animals), for instance we don't write a book about the subject we don't know about.

There is no empirical evidence that our brain tries to create content which is a short distance from content (books, conversations, etc.) created under exceptional circumstances (when doubts are minimal). When we are driving, and we do not know what is in front of us, we usually just slow down or stop the car. But what content defines "not knowing"? Is there empirical evidence about the unknown? The unknown can only be an abstract concept, expressed through probability theory or a logical equivalent. Is there empirical evidence that probabilities are reducible, that there is a simpler logical equivalent? No, quite the opposite.

The only trade-off seems to be between costs (time complexity, etc.) and understanding/control. A prior measure which we understand and/or control may mean much more costs than an emergent (thus, inconsistent and uncontrollable) prior measure which just minimizes some distance. But this trade-off is not new, and it is already present in all industries which deal with some safety risk (which is essentially all industries). Distances are efficient for proof of concepts (pilot projects), when the goal is to show that we are a short distance from where we want to be. But safety (as most features) is not being at a short distance from being safe[4]. "We were at a short distance from avoiding nuclear annihilation" is completely different from "we avoided nuclear annihilation". To avoid nuclear annihilation we need (probability) measures, not only distances.

[4]See for instance https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec

5. Empiricism also implies that deep learning is a random sampling method

Even when ignoring all existing theories about Statistics, logical consistency implies that deep learning is a random sampling method.

Deep learning is based on data and thus can be always be mathematically defined as a part of Bayesian inference. The only doubt is whether such definition is useful or not. Then, there are no non-informative priors, thus bias or prior (that is, some theory) is unavoidable.

Note that despite the theory being unavoidable, humans may not need to understand it, for instance if it results in more power to produce new human-understandable theories[48][49]. After all, all theories are temporary. Non-understandable human theories are inherently non-scalable, because they cannot be trusted, but can be used to produce trustable, scalable things, such as trustable, scalable theories. In the following we will just assume that some theory is unavoidable, whether or not it is human-understandable.

This is today the mainstream approach to Deep Learning. The problem is that such tools that would empower us to produce new human-understandable theories, did not exist prior to Deep Learning (and we could argue that they do not exist yet, but that is not needed for the purposes of this article). Thus, using Deep Learning, however we define it, we created (or will create) such new tools, which could only be created by trial-and-error (experiments which produce empirical evidence). But trial-and-error in the context of Bayesian inference (where we are, because we use data) is just a random sampling method.

We could now argue that Deep Learning includes a random sampling method, but it is more than that. However, whatever it is more than that, it was created by humans, and it will be replaced, sooner or later, by a creation developed using the random sampling method included in Deep Learning itself. In conclusion, for the same reasons that we can ignore all existing theories of Statistics, we can also ignore everything in Deep Learning except its random sampling method. For the sake of logical consistency.

Note that there is also the logical possibility that Deep Learning will not give us more power to produce new human-understandable theories, thus we cannot ignore all existing theories about Statistics. But this still leads to the same conclusion that Deep Learning is a random sampling method, as discussed in the previous sections.

6. Application: Incentives in a representative democracy vs. alignment

The fact that both deep learning and human learning are random sampling methods has consequences for the alignment problem[50]. Because in most sufficiently complex systems (certainly in a system with the goal of solving many and different complex problems), it is extremely unlikely to find a sufficiently simple statistical metric that it is not misleading (in other words, that it is not rogue). This is not a problem but a feature, since we can define probabilities as the absence of information, unpredictability, thus, if it would not be likely misleading it would not be random/unpredictable and if it would not be random it would not generalize well (by definition of generalization in a Bayesian context). For instance, the economic growth of any country (however we measure it) is a misleading statistical metric.

Another example: in a representative democracy there are incentives in place with the theoretical goal of aligning the behaviour of elected officials (who are much more powerful than the average citizen) to serve the best interests of the average citizens.

While most elected officials serve at least some of the interests of the average citizens, some dosis of anarchy (however small) is always present in (and we could argue that it is essential for the stability of) a representative democracy. By anarchy we mean, some citizens (including elected officials) disrespect the laws and decieve each other in unpredictable ways.

In fact, a representative democracy is an extremely inneficient way of governing. The only reason that many citizens still want it is that it empowers the average citizen at the expence of a rulling elite with less power. Why is that a good thing? Because in many thousands of years, no society could properly "align" a ruling elite to serve the interests of the average citizen, thus the ruling elite should have enough power to prevent complete anarchy, but not much more than that.

Note that it is not a matter of how smart or powerful the rulling elite is. In fact, the rulling elite must also manage natural resources such as rocks (diamonds for instance), which are literally "dumb as a rock" and they never did it efficiently. So, it is extremely unlikely that a "super-human" ruling elite (even if equipped with artificial intelligence with infinite scale, since the intelligence ratio of a human against a rock is also infinite) could somehow be aligned to serve the interests of the average citizen. Certainly, there exist many good solutions for such alignment problem, however the good solutions are a tiny fraction of the possible solutions. Since deep learning is a random sampling method (as human learning), it is extremely unlikely that will select one of the good solutions.

We address now consequences to weak-to-strong generalization[50]. Weak-to-strong generalization is about using a weaker model (which we trust) to "elicit" knowledge from a stronger model (which we

do not trust). The authors themselves acknowledge that this just a (far from perfect) analogy with the alignment problem, that they "hope" may lead to progress. About the "hope" we cannot say much. About the analogy we can say that it doesn't change our findings, because a "weak model eliciting knowledge from a strong" model is after all still some complex enough Bayesian model whose "good generalization" properties still come from unpredictability/randomness: either the model as a whole (that is, its weak+strong parts) generalizes well and it is unpredictable, or it can be trusted on and it does not generalize well.

The way out is to split a complex system into simple systems, if we can do that, then some simple statistical measures become trustworthy. We will explore such possibility in the next section.

7. Application: Alignment for fully autonomous machines

Neural Networks and symbolic AI do not solve by themselves the alignment problem for fully autonomous machines. A well known solution is twofold: 1) the laws (constraints) are ultimately "interpreted" by judges (who are human beings) who have the authority to explore all its ambiguities to obtain the concrete constraint they think it is best for the concrete case they are judging; 2) this necessarily generates some unpredictability, the costs of it are on the person who looses the judicial case, due to the principle "Ignorance of law excuses no one". This means that people must act cautiously and only take actions that they are confident most judges will interpret as lawful. The degree of caution is adapted to the circumstances, for instance if there is some medical emergency they will be less cautious than if there is a conflict due to spurious reasons, and this also affects the interpretation of the law by the judges.

Thus, the degree of confidence (or equivalently a probability) on predictions about the consequences of the decisions is an ingredient of many decisions people make. This requires a probabilistic model of the real world where the person making the decision is included in the model (we could argue that this is consciousness, but such discussion is beyond the scope of this article).

A model of the real world involves time dynamics, the most general such model is a quantum model. But generality by itself could not help when there are infinite variables. Crucially, choosing a prior probability distribution where the quantum Hamiltonian (that is, the generator of the time-evolution) is a bounded operator is consistent with the real-world, because the available energy to a person or robot is finite (in most cases). This allows us to use few variables for a small enough energy interval (using the Krylov subspace, and wavelets to distinguish energy intervals and also energy resolutions), for a specific reason (finite energy interval and/or resolution) related to the specific problem we are solving (predicting real consequences).

Note that in many cases, the energy interval may be too large and additional assumptions are

required. In some cases, the additional assumptions may turn the assumption of the finite energy interval irrelevant, then the use of a quantum model may not be advantageous. But in many other cases, predicting the consequences of a decision is mostly about physics. Moreover, concepts unrelated to physics can nevertheless be grounded in (defined from) examples from physics, specially when the model of the real world is Turing complete (that is, it can be used to solve any other problem).

Many authors argue that all these features can "emerge" in a sufficiently large Neural Network using enough training data. But a relevant question is: is there a differentiable symbolic AI where all these features are explicit? Yes, the mathematical formalism of Quantum Mechanics has all these features explicitly: it includes probabilities explicitly, it is linear (thus differentiable) and it allows reducing a large problem into a sum of many problems with few variables each (using the spectral theorem and the Krylov subspace for continuous spectra). Moreover, it includes a uniform measure for an infinite-dimensional sphere, so it can be used to define many infinite-dimensional mathematical problems and also its approximate numerical solution, thus the approximations are better understood.

8. Application: Previous knowledge as step-by-step learning

There are many ways of including the previous knowledge into the prior probability distribution[51][52], including just changing the size of the sample space to allow learning ambiguous intermediate concepts.

For instance, classifying if a statement is a fact or fiction has some ambiguity, but most people agree on whether or not many statements are a fact or fiction (with many exceptions, for sure). When we teach a child to not lie, the child understands that it does not mean to never lie, under any circumstance. Thus, we can always include a pre-trained classifier of facts/fiction, whose output will be added to the original input, forming the input of another Bayesian model. Then the Bayesian model has access to the previous knowledge included in the classifier, but it still has the freedom to completely ignore it in extreme cases, since it also has access to the original input.

9. Application: Bayesian inference for non-deterministic functions

Most real-world functions are non-deterministic. For instance, no one looks to a photo of a real gorilla and says with absolute certainty that it is not a man in a gorilla suit. This ambiguity has major advantages in a dynamic world where new information may always appear. Deep learning using neural networks is only good to fit deterministic functions (including the mean-values of non-deterministic functions), but it fails badly in most real-world functions[42] since it cannot fit

the many relevant correlations of subsets (not just pairs) of non-deterministic outputs.

Neural networks have this feature by construction, since the likelihood of the outputs is assumed uncorrelated to allow for the definition of a loss function to be minimized without predicting the variance of the outputs. One can always promote a probability distribution to be the deterministic function we want to study, but that is inefficient since then the likelihood of the outputs (or a value that encodes it[40]) which stores a large amount of information, becomes redundant. A better alternative would be to consider the most general statistical model, which implies considering non-deterministic functions, and it predicts the likelihood of the outputs[16], including correlations and variances; this includes, but it is more general than considering arbitrary prior knowledge for the parameters of a specific statistical model[1]. Note that it is not due to a lack of computational power since modern neural networks can fit very complex deterministic functions and fail badly[42][43] in relatively simple non-deterministic functions (e.g. catastrophic forgetting or the need of calibration to have some probabilistic guarantees[43]).

Although there are many attempts to extend neural networks[42][1][43][45], we can easily see that in general, there are many more relevant correlations of subsets (not just pairs) of outputs than there are outputs. Therefore, just the fact that neural networks are good to fit the mean-values of the outputs is no indication that they are a good starting point to find a method to fit all the relevant correlations. We better take a step back and use everything we already know.

And what do we know about deep neural networks? Not so much, except the fact that the present evidence suggests that they are black boxes[42][45]. It does not look a great idea to extend a black box, at least without further information. And what do we already know about correlation functions and the necessary ambiguity to adapt to new information ? A lot and since a long time, starting with Bayesian inference:

> *"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth? We know that he did not come through the door, the window, or the chimney. We also know that he could not have been concealed in the room, as there is no concealment possible. When, then, did he come?"*
>
> — Sherlock Holmes in The Sign of the Four (1890)

Bayesian inference doesn't look so hard, does it? Moreover, since neural networks can in principle be used to fit any function including probability distributions, they seem to be compatible with Bayesian inference as long as we know what should be fitted. In fact, supervised learning can be defined as a particular case of reinforcement learning[53]. In reinforcement learning, deep neural networks can be used, and any reward can be defined, including a reward which does not assume that the likelihood of the outputs is uncorrelated. And incorporating Bayesian inference in Reinforcement Learning

provides a machinery to action-selection under uncertainty and to incorporate prior knowledge into the algorithms[54].

Functions are in general infinite-dimensional spaces, so it makes sense to look for measures in infinite-dimensional spaces. While the Lebesgue measure cannot be defined in a Euclidean-like infinite-dimensional space[55][56], it is well known since many decades that a uniform (Lebesgue-like) measure of an infinite-dimensional sphere can be defined using the Gaussian measure and the Fock-space (the Fock-space is a separable Hilbert space used in the second quantization of free quantum fields)[47]. Such a space can parametrize (we call it the free field parametrization) the probability distribution of another probability distribution with sample space given by the direct product of the base space and the field (say $\mathbb{R}^{n+1}$). A probability distribution with sample space given by the direct product of the base space and the field can model a non-deterministic function since we can always choose a measure (the Gaussian measure for instance) which covers the whole base space, with the non-deterministic function evaluated at each point of the base-space given by the conditional probability distribution conditioned on each point of the base-space. Note that due to the relation between Fock-spaces and tensor products, we can consider regions of the sample space as small as we want and there is null measure for a null marginal probability in each one of these regions, and for each component of the Fourier series corresponding to such region; thus the conditional probability distribution conditioned on each point of the base-space is well-defined everywhere except in sets with null measure.

Thus, the free field parametrization is appropriate for non-deterministic functions but not for deterministic functions, since the prior (uniform measure of an infinite-dimensional sphere, we call it the uniform prior from now on) attributes null measure to deterministic functions. If we were interested in deterministic functions, we would be in trouble[55][56]. Now that we found a probability space and a candidate prior, we need to look for other candidate priors to choose from.

In the free field parametrization, the uniform prior defines a vector of the Hilbert space which when used as the prior for Bayesian inference with arbitrary data generates an orthogonal basis for the whole Fock-space. Such basis is related with a point process, with the number of points with a given feature corresponding to the number of modifications to the uniform prior (in the part of the sample space corresponding to such feature). Since Bayesian inference with any other prior can be seen as a combination of the results of different Bayesian inferences with the uniform prior for different data (eventually an infinite amount of data for the cases with null measure), then the uniform prior in the free field parametrization is appropriate for non-deterministic functions in the absence of any other information.

10. Application: Digital-first Mathematics and Mathematical Modelling

Most mathematics and mathematical modelling are human-first (or analog-first or book-first). That is, the framework is optimized such that the computers are dispensable, computers play a mere productivity role. For instance, a mathematics book is such that it could be written by hand, but by writing it in the computer it can reach students faster and cheaper. The same happens with most mathematical models of complex systems: they can in principle be written by hand, the computer just increases productivity.

But in the next few years, mathematics and mathematical modelling will become digital-first. That is, the framework is optimized such that the human interaction is dispensable, humans play a mere auditing role. For instance, we could have new alternatives to complex analysis that are born digital. As it happens today with digital photos, humans are provided with an analog view (in a screen or on paper) of the photo and check if they like it, but the photo itself is digital, and it has more information/resolution than what humans can easily distinguish. As it is progressively happening with autonomous driving, where humans audit the car's autonomous progress but don't effectively drive as much.

Bayesian inference (or approximations to it, such as deep neural networks) is not deterministic. Therefore, Machine Learning can create deterministic mathematics, but it is not deterministic mathematics itself. Thus, we still need some kind of mathematics. All current Computer Algebra Systems and Proof assistants were optimized for a human-first Mathematics. They are computationally inefficient, verbose, logically inconsistent, designed for a human-first mathematics and to please the human user.

An alternative is the Haskell library Egison (in its sweet-egison implementation[57], a similar alternative would be Metamath[58]). Haskell is a pure functional language (thus easy to run in parallel architectures, see www.acceleratehs.org) and it is the only pure functional language that is often as fast as C. Moreover, we can easily include C code in Haskell in it in the exceptional cases where C is much faster. Then the Haskell's Egison library allows us to implement (probably) the most expressive and advanced Computer Algebra System. Thus, we have expressiveness, flexibility and speed, despite that it requires a lot of human expertise to do Mathematics from scratch with Haskell's Egison (which is ok for Digital-first Mathematics). Since the human user plays an auditing role, another language such as Python would not be much more useful than Haskell (a complex Python function is also hard to audit by a non-expert, moreover machine learning knows how to write Haskell interfaces to the Python libraries that could eventually be needed) and we don't need many human auditors in Mathematics (due to its Universality) in the World, these few people can

be specialists in Haskell, Egison and Mathematics.

Most human users will see/edit views of the Haskell code (constraints /explanations in plain English for instance), but not the code itself. These views are non-deterministic and can be a source of errors, which are ultimately audited by specialists in Haskell and Mathematics.

11. Application: Learning interesting questions, for which we can find a predictive model

Perhaps the biggest advantage of defining deep learning as a random sampling method in Bayesian inference, is that it allows us to apply deep learning (or alternative sampling methods) to more abstract problems than (eventually unsupervised) machine learning.

We all love to make a prediction and to be right ("I told you"). Thus, we tend to try to find questions/problems for which we can learn the answer to. Note that, our memory is limited, so this is one relevant way (perhaps even the primary way) that allows us to select the information to save in memory, since the more predictive a model is, the less memory it requires[59]. Since we cannot answer everything anyway, we try to find those questions we can answer to, to allocate scarce resources to them.

So we have a hierarchical inference problem, where the computation of the posterior probability includes another inference problem. And the hierarchy can have as many levels as we wish. But there is only one human brain or only one machine, so it does not make much sense to say that an inference problem in an intermediate level of the hierarchy is "machine learning" or a "neural network" analogous to the brain. However, once we define deep learning as a random sampling method in Bayesian inference, then it makes perfect sense to apply it to the intermediate levels of the hierarchy, since a hierarchy of beliefs/probabilities is a natural concept.

References

[1] Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., & Hutter, F. (2022). Transformers Can Do Bayesian Inference. *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=KSugKcbNf9

[2] Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, *1*(1), 1–40. https://doi.org/10.1214/06-BA101

[3] Harrison, W. J., Bays, P. M., & Rideaux, R. (2023). Neural tuning instantiates prior expectations in the human visual system. *Nature Communications*, *14*(1), 5320. https://doi.org/10.1038/s41467-023-41027-w

[4] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 1050–1059. New York, NY, USA: JMLR.org. https://doi.org/10.5555/3045390.3045502

[5] Hermann, J., Spencer, J., Choo, K., Mezzacapo, A., Foulkes, W. M. C., Pfau, D., … Noé, F. (2023). Ab initio quantum chemistry with neural-network wavefunctions. *Nature Reviews Chemistry*, 1–18. https://doi.org/10.48550/arXiv.2208.12590

[6] LeCun, Y. (2019). *The Epistemology of Deep Learning - Videos | Institute for Advanced Study*. Retrieved from https://www.ias.edu/video/DeepLearningConf/2019-0222-YannLeCun

[7] Eaton, M. L., & Freedman, D. A. (2004). Dutch book against some 'objective' priors. *Bernoulli*, *10*(5), 861–872. https://doi.org/10.3150/bj/1099579159

[8] Ngo, R., Chan, L., & Mindermann, S. (2022). *The alignment problem from a deep learning perspective*. arXiv. https://doi.org/10.48550/arXiv.2209.00626

[9] Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2023). *Fundamental Limitations of Alignment in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2304.11082

[10] CAIS. (2023). *Statement on AI Risk*. Retrieved from https://www.safe.ai/statement-on-ai-risk#open-letter

[11] Pinker, S. (1994). *The language instinct*. New York, NY, US: William Morrow & Co. pp190-191.

[12] Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press. pp15-16.

[13] Rasmussen, C. E., Williams, C. K. I., Press, M. I. T., Bach, F., & (Firm), P. (2006). *Gaussian Processes for Machine Learning*. MIT Press. Retrieved from http://gaussianprocess.org/gpml/chapters/RW.pdf

[14] Laumann, F. (2020). When machine learning meets complexity: why Bayesian deep learning is unavoidable. Retrieved from https://towardsdatascience.com/when-machine-learning-meets-complexity-why-bayesian-deep-learning-is-unavoidable-55c97aa2a9cc

[15] Jihan, N. (2019). *Re: How does Bayesian inference compare against other machine learning models?* Retrieved from https://www.researchgate.net/post/How-does-Bayesian-inference-compare-against-other-machine-learning-models/5d0a1ad8d7141b7d8643d972/citation/download

[16] Dupont, E., Kim, H., Eslami, S. M. A., Rezende, D. J., & Rosenbaum, D. (2022). From data to functa: Your data point is a function and you can treat it like one. *Proceedings of the 39th International Conference on Machine Learning*, 5694–5725. PMLR. Retrieved from https://proceedings.mlr.press/v162/dupont22a.html

[17] Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. Retrieved from http://cogprints.org/3106

[18] Roychowdhury, S., Diligenti, M., & Gori, M. (2021). Regularizing deep networks with prior knowledge: A constraint-based approach. *Knowledge-Based Systems*, *222*, 106989. https://doi.org/10.1016/j.knosys.2021.106989

[19] Turner, R., Hung, J., Frank, E., Saatchi, Y., & Yosinski, J. (2019). Metropolis-hastings generative adversarial networks. *International Conference on Machine Learning*, 6345–6353. PMLR. https://doi.org/10.48550/ARXIV.1811.11357

[20] Kaplanis, C., Shanahan, M., & Clopath, C. (2018). Continual reinforcement learning with complex synapses. *International Conference on Machine Learning*, 2497–2506. PMLR. https://doi.org/10.48550/arXiv.1802.07239

[21] Beskos, A., Roberts, G., Stuart, A., & Voss, J. (2008). MCMC methods for diffusion bridges. *Stochastics and Dynamics*, *8*(03), 319–350. Retrieved from https://authors.library.caltech.edu/69496/1/stuart74.pdf

[22] Vollmer, S. J. (2015). Dimension-Independent MCMC Sampling for Inverse Problems with Non-Gaussian Priors. *SIAM/ASA J. Uncertain. Quantification*, *3*, 535–561. https://doi.org/10.48550/arXiv.1302.2213

[23] Chen, V., Dunlop, M. M., Papaspiliopoulos, O., & Stuart, A. M. (2018). *Dimension-Robust MCMC in Bayesian Inverse Problems*. arXiv. https://doi.org/10.48550/ARXIV.1803.03344

[24] Cui, T., Tong, X. T., & Zahm, O. (2022). Prior normalization for certified likelihood-informed subspace detection of Bayesian inverse problems. *Inverse Problems*, *38*(12), 124002. https://doi.org/10.48550/ARXIV.2202.00074

[25] Villa, U., Petra, N., & Ghattas, O. (2021). HIPPYlib: An Extensible Software Framework for Large-Scale Inverse Problems Governed by PDEs: Part I: Deterministic Inversion and Linearized Bayesian Inference. *ACM Trans. Math. Softw.*, *47*(2). https://doi.org/10.1145/3428447

[26] Beskos, A., Girolami, M., Lan, S., Farrell, P. E., & Stuart, A. M. (2017). Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, *335*, 327–351. https://doi.org/10.1016/j.jcp.2016.12.041

[27] Coullon, J., & Webber, R. J. (2021). Ensemble sampler for infinite-dimensional inverse problems. *Statistics and Computing*, *31*(3), 28. https://doi.org/10.48550/arXiv.2010.15181

[28] Carreira, J., Koppula, S., Zoran, D., Recasens, A., Ionescu, C., Henaff, O., ... Jaegle, A. (2022). *HiP: Hierarchical Perceiver*. https://doi.org/10.48550/arXiv.2202.10890

[29] Benjamin, A. S., & Kording, K. (2018). *Improving generalization by regularizing in L^2 function space*. Retrieved from https://openreview.net/forum?id=H1l8sz-AW

[30] Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *Acta Numerica*, *30*, 87–201. https://doi.org/10.1017/S0962492921000027

[31] Alavisamani, N., & Aghababa, H. (2018). *Application of Quantum Gradient Descent as a Learning Algorithm for Factorization Machines: Quantum Artificial Intelligence*. https://doi.org/10.1145/3200947.3201025

[32] Singh, C. (2022). *Useful interpretability for real-world machine learning*. EECS Department: University of California at Berkeley, Berkeley, California. Retrieved from https://digicoll.lib.berkeley.edu/record/264283

[33] Zhang, Z., Hamadi, H. A., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access*, *10*, 93104–93139. https://doi.org/10.1109/ACCESS.2022.3204051

[34] López-Pastor, V., & Marquardt, F. (2023). Self-Learning Machines Based on Hamiltonian Echo Backpropagation. *Physical Review X*, *13*(3), 031020. https://doi.org/10.1103/PhysRevX.13.031020

[35] Lu, Y., & Lu, J. (2020). A universal approximation theorem of deep neural networks for expressing probability distributions. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 3094–3105. Red Hook, NY, USA: Curran Associates Inc. https://doi.org/10.5555/3495724.3495984

[36] Kratsios, A., Zamanlooy, B., Liu, T., & Dokmanić, I. (2021). Universal Approximation Under Constraints is Possible with Transformers. *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=JGO8CvG5S9

[37] Pérez, J., Marinković, J., & Barceló, P. (2019). On the Turing Completeness of Modern Neural Network Architectures. *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=HyGBdo0qFm

[38] Hernández, A., Millerioux, G., & Amigó, J. M. (2022). *Differentiable programming: Generalization, characterization and limitations of deep learning*. arXiv. https://doi.org/10.48550/arXiv.2205.06898

[39] Hennig, P., Osborne, M. A., & Kersting, H. P. (2022). *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press. Retrieved from https://www.probabilistic-numerics.org/assets/ProbabilisticNumerics.pdf

[40] Bojun, H., & Yuan, F. (2023). *Utility-Probability Duality of Neural Networks*. arXiv. https://doi.org/10.48550/arXiv.2305.14859

[41] Novak, E., & Wozniakowski, H. (2008). Tractability of Multivariate Problems, Volume I: Linear Information, European Math. *Soc., Zürich*, *2*(3).

[42] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, *30*. Retrieved from https://proceedings.neurips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf

[43] de Grancey, F., Adam, J.-L., Alecu, L., Gerchinovitz, S., Mamalet, F., & Vigouroux, D. (2022). Object Detection With Probabilistic Guarantees. *Fifth International Workshop on Artificial Intelligence Safety Engineering (WAISE 2022)*. München, Germany. Retrieved from https://hal.archives-ouvertes.fr/hal-03769683

[44] Li, Y., & Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in Neural Information Processing Systems*, *31*. Retrieved from https://proceedings.neurips.cc/paper/8038-learning-overparameterized-neural-networks-via-stochastic-gradient-descent-on-structured-data.pdf

[45] Egele, R., Maulik, R., Raghavan, K., Lusch, B., Guyon, I., & Balaprakash, P. (2022). Autodeuq: Automated deep ensemble with uncertainty quantification. *2022 26th International Conference on Pattern Recognition (ICPR)*, 1908–1914. IEEE. https://doi.org/10.48550/ARXIV.2110.13511

[46] Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. Retrieved from https://cognitivemedium.com/magic_paper/assets/Hornik.pdf

[47] Peterson, A. (2019). *Gaussian Limits and Polynomials on High Dimensional Spheres* (PhD Thesis, University of Connecticut). University of Connecticut, Storrs, CT USA. Retrieved from https://opencommons.uconn.edu/dissertations/2137

[48] Steinhardt, J. (2023). Language Models as Statisticians, and as Adapted Organisms. Retrieved from https://simons.berkeley.edu/talks/jacob-steinhardt-uc-berkeley-2023-08-16

[49] Chen, Y., Zhong, R., Ri, N., Zhao, C., He, H., Steinhardt, J., … McKeown, K. (2023). *Do Models Explain Themselves? Counterfactual Simulatability of Natural Language Explanations*. arXiv. https://doi.org/10.48550/arXiv.2307.08678

[50] Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., … Wu, J. (2023). *Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision*. arXiv. https://doi.org/10.48550/arXiv.2312.09390

[51] Jin, X., Zhang, D., Zhu, H., Xiao, W., Li, S.-W., Wei, X., … Ren, X. (2022). *Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora*. NAACL 2022. https://doi.org/10.48550/arXiv.2110.08534

[52] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., … Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, *5*(3), 220–235. https://doi.org/10.1038/s42256-023-00626-4

[53] Barto, A. G., & Dietterich, T. G. (2004). Reinforcement Learning and Its Relationship to Supervised Learning. *Handbook of Learning and Approximate Dynamic Programming*, 47–64. Retrieved from https://web.engr.oregonstate.edu/~tgd/publications/Barto-Dietterich-03.pdf

[54] Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian Reinforcement Learning: A Survey. *Foundations and Trends® in Machine Learning*, *8*(5-6), 359–483. https://doi.org/10.48550/arXiv.1609.04436

[55] Baker, R. (1991). "Lebesgue measure" on R^∞. *Proceedings of the American Mathematical Society*, *113*(4), 1023–1029. https://doi.org/10.1090/S0002-9939-1991-1062827-X

[56] Baker, R. (2004). "Lebesgue measure" on R^∞, II. *Proceedings of the American Mathematical Society*, *132*(9), 2577–2591. https://doi.org/10.1090/S0002-9939-04-07372-1

[57] Egi, S., & Nishiwaki, Y. (2020). Functional Programming in Pattern-Match-Oriented Programming Style. *The Art, Science, and Engineering of Programming*, *4*(3), 7:1–7:32. https://doi.org/10.22152/programming-journal.org/2020/4/7

[58] Polu, S., & Sutskever, I. (2020). *Generative Language Modeling for Automated Theorem Proving*. arXiv. https://doi.org/10.48550/arXiv.2009.03393

[59] Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., … Veness, J. (2023). *Language Modeling Is Compression*. arXiv. https://doi.org/10.48550/arXiv.2309.10668