PUBLIC SUBMISSION

Received: May 07, 2025
Tracking No. mae-o7l3-scbg
Comments Due: May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0065
Comment on FR Doc # 2025-07332

Submitter Information

Name: Nicholas Teague

General Comment

Please see letter attached

Attachments

Feedback_towards_development_of_a_2025_national_Artificial_Intelligence

Feedback towards development of a 2025 national Artificial Intelligence R&D Strategic Plan

Nicholas Teague
Nicholas Teague LLC

In response to
Docket ID No. NSF-2025-OGC-0001

https://www.federalregister.gov/documents/2025/04/29/2025-07332/request-for-information-on-the-development-of-a-2025-national-artificial-intelligence-ai-research?et_rid=1113557638&et_cid=5607774

The RFI requests "how the previous administration's National Artificial Intelligence Research and Development Strategic Plan (2023 Update) can be rewritten so that the United States can secure its position as the unrivaled world leader in artificial intelligence by performing R&D to accelerate AI-driven innovation, enhance U.S. economic and national security, promote human flourishing, and maintain the United States' dominance in AI while focusing on the Federal government's unique role in AI research and development (R&D) over the next 3 to 5 years"

Response:

This letter won't go into specific drafting language or guidelines for the strategic plan, rather I would like to draw from my own experience in the field to convey some matters of importance that the prior administration did not address fully in a public setting, I expect at least partly because of the pace of change in the field has been so rapid.

Speaking as someone who has followed the AI research domain closely for several years, my impression of US research is that although it is recognized that the vast majority of artificial intelligence algorithms are probabilistic by nature, the extent to which the probabilistic nature of inputs and outputs to an algorithm are considered in a fine grained manner in the literature is surprisingly low. In earlier conventions of generative models, before the advent of large language models, the topics of probabilistic sampling to algorithms were most commonly discussed when sampling from gaussian or related source distributions and related hyperparameters in model training, such as those used to seed random generated images to a variational autoencoder or generative adversarial network. The output of various other machine learning conventiond mainly considered probabilistic components from the standpoint of uncertainty quantification, such as to the result of a classification evaluation or to derive error curves surrounding a regression forecast.

As the field has progressed to the large language model conventions using transformer architectures and etc, the probabilistic settings to common forms of consumer language API's have mostly been limited to a "temperature setting", which is intended to adjust for how predictable with the output of a generated token, which is often used as a proxy for "creativity" or "variability" in a language model setting.

Under the hood, these language models in modern conventions have begun to establish new forms of probabilistic sampling into "next token prediction" operations (aka speculative decoding), which may have benefits in comparison to an exact derivation of most likely token predictions related to latency or enlarged context window of inference in comparison to fully deriving the exact solution as a single most probable next token operation.

In the image modality generative model space, a newer convention called diffusion models that are now in mainstream use is even more natively probabilistic, as the pixels / color / and shadings for generated images are progressively populated through a series of learned probabilistic model samples.

The native leveraging of probabilistic components in the diffusion model setting possibly creates an unprecedented impasse towards mainstream AI research that in many cases has in recent years been been predominantly published in open settings with research shared in public venues such as arXiv. In most cases of preceding work in the AI field, drawing samples from probabilistic distributions has most often been conducted far "under the hood" of operating systems and hardware, leveraging different channels of research that was often published in different venues than AI. Although there have been public (if not well known) python libraries like Edward (which later became TensorFlow Probability) or PyTorch's similar Pyro library that were commonly used for deriving or specifying more intricate forms of probability distributions for applications like variational autoencoders, in most cases the randomness itself was still sampled in a common way to other algorithms by way of leveraging those sources of randomness available to mainstream operating systems and hardware.

Those sources of randomness available to mainstream operating systems for the most part have traditionally relied on forms of algorithms known as "pseudo random number generators", which typically take some kind of "seed" of randomness from eg a clock time stamp or operating system state, and then extract that to a large integer that has properties resembling randomness. It is an unfortunate misconception that such pseudo random generators are themselves truly random. Various conventions each have their own patterns and edge cases, for example a pseudo random generator may cycle through the set of generated outputs and thus be repeating common samples for common seeds (attempting to describe in a simplified way), or the source of seeding may have some form of pattern or repetition that can likewise be extracted. The implications of "non-certifiable randomness" are profound, and can have implications towards sensitive settings like encryption, cryptography, or related matters.

[Consider as speculation that with the prior administration's call for industry to recognize and adapt to the reality that many common programming languages, including some that are themselves the bedrock of modern operating system builds (like C, C++, etc), have inherent memory exposure channels that are only circumsvented by replacing with alternative memory safe programming languages like Rust or otherwise much slower langauges like Python, we can expect that when you couple issues with randomness backdoors and issues with programming lagnauge backdoors, there ****omitted]

The obvious followup question then becomes how to get access to true and "certifiable randomness" in an algorithmic context, where we not only need to know that these certifiable properties hold, but we need to access in a near instantaneous manner (as the chips themselves are operating at ~nanosecond time cycles for transistor operations and then the software layer based on context can be expected to be operating at a slightly slower pace). One may look to more pure sources of randomness in our environment, this itself is a complicated question though.

A different channel of research that is realizing comparable pace of progress to the AI field is less widely followed in mainstream media, associated with the conventions of quantum computing. Although some quantum circuits target operations that are deterministic in nature, the properties of quantum circuits are inherently stochastic due to qubit and qubit gate noise properties, what is commonly considered the fundamental obstacle to truly scalable quantum circuits approaching the scales that are expected to lead to those types of applications that brought so much attention to the field in the first place (such as cryptography and encryption breaking algorithms that will eventually be able to circumvent encryption preceding those recently released recommended standards from the US Department of Commerce's NIST institution associated with post quantum cryptography).

The quantum computing field is somewhat hard to wrap your head around at times (is an understatement). Part of the challenge with dependability of various conventions in quantum computing is that quantum circuits are not just a product of their settings, they are potentially exposed to influence of other quantum circuits, even those sourced on different hardware, solely due to the presence of some joint sources of quantum entanglement (eg for sensors whose input are fed into different circuits of overlapping scope). For fully deterministic algorithms this may not be a complicating factor when the circuits are successfully implemented without breaking down through environmental interference, but for those probabilistic algorithms that quantum circuits are so adept at we must become at least aware to the potential influence of other circuits. On it's surface this presents enormous security concerns, but as the reality of the world is that quanutm computers are being made available now on nearly every continent, it is a reality that we are going to need to find a way to work around.

Thus, a fundamental challenge facing the field of artificial intelligence is that as our AI research proceeds deeper into the realm of probabilistic algorithmic components, such as those that are a part of modern paradigms of diffusion and transformer architectures often adapted towards image and

language generative applications, the true fundamental innovations in the field going forward are likely going to be dependent not just on traditional back-propagation methods that GPU hardware has shown to be so adept at implementing on parallelized systems (that are now being scaled in epic proportions), they are going to need even more innovation at the intersection of AI and probabilistic algorithms that quantum computing is so adept at. This means we are going to need to re-evaluate what are the boundaries of research distribution in the public space and how we can grow talent in emerging tech even with potential of non-isolated circuitry.

This researcher has attempted to reason through a few of these matters, and would like to propose that one natural boundary for how research and industry can be kept tractable and sane in this environment is to make the blanket requirement that the only settings where fully deterministic algorithms are suitable (such as those that would be enabled by the eventual development of fully fault tolerant quantum circuits or otherwise by modern paradigms of quantum simulators that are capable of being implemented on classical hardware such as GPU clusters), would be to limit accessibility of fully deterministic probabilistic algorithms to the defense sector.

If we as a nation are to remain competitive in this international economy, we will still need to increasingly adapt emerging quantum computing capabilities not just into defense sector, but also into mainstream commercial industry. To achieve that, we will need to find a way to _prioritize_ between a hierarchy of potentially divergent applications with potentially divergent objectives / performance metrics. A few ways we could achieve that could include:

- for cloud sourced samples from quantum circuits, we could auction off access to these resources based on access to shorter time lags (or shorter time lag distributions) between the circuit operation and the cloud sourced sample, such that prioritization would be achieved by ensuring the most economically significant applications would be capable of operation, while at the same time those shortest time lags or otherwise specific reserved time lag distributions characteristics could be reserved for those mission critical or public interest applications. I expect that such auctions could even be structured in a manner to funnel a new form of tax revenue to local/state/federal institutions in a manner resembling the FCC auctions of wireless spectrum.
- For those applications in our economy that aren't directly incorporating quantum circuits into their own formulation, we could still channel the influence of quantum circuits by way of sampling purely eg gaussian noise sampled from quantum circuits into inference operations, such that models could be trained to account for such noise being present in inference except for settings where the application overlaps with some higher priority quantum computation, which would gain ability to stochastically influence the classical algorithm in direction of weighted probabilities of the higher priority system.
    - (It is not something I have fully been able to formulate with provable theoretical backing, but I have some expectation that as the scale of quantum sampled entropy by noise injections to classical circuits reaches a sufficiently wide scale in the macro setting, **this form of scaled noise can become an alternative to fully isolated circuits and settings**, in a manner resembling how with increasing number of parameters neural networks experience the "double descent" phase change towards the convergence to generalized models that were one of key enabling factors for modern paradigms of learning. Aspects

of this paragrphs were influenced by some preprints that I had shared to arXiv including the work "Stochastic Perturbations of Tabular Features for Non-Deterministic Inference with Automunge" as well as the paper "Geometric Regularization from Overparameterization". I found some supplemental supporting evidence by way of an ICLR 2024 conference talk by a Genentech researcher that I previously blogged about.)

- For those more sophisticated systems that are directly built on top of custom designed quantum circuits in some fashion, we can allocate circuits with error distributions appropriate to the domain.
  - For example, for "batch" type evaluations, it may be sufficient to take advantage of "NISQ" hardware ("noisy intermediate scale quantum circuits), which although are noisy in isolation, with sufficient repetition they can be expected to hone in on solutions - these will still be suitable for researchers in an academic or educational setting as well.
  - For streaming applications that require continuous samples, and particularly for those that are exposed to mission critical or safety related domains, the conventions of quantum annealing (see eg the work of D-Wave Quantum who recently published a Science journal paper validating their platform), for although their form of optimization algorithms are also sampling from distributions, the distribution of sampled solution quality can be expected as much "thinner tailed" than gate model quantum circuits that are attempting to apply variational quantum circuits (eg the "QAOA" circuit) to this application. One way to think about it is that although a gate quantum circuit is exposed to at least the remote possibility of returning a fully random output, a quantum annealer is only exposed to the tail of the Boltzmann distribution for returned solution quality, which even when not returning a "global minima" optimal solution, can be expected to at least return samples of approximately optimal solutions. This dynamic can be expected as especially beneficial in settings with the tightest latency requirements.

Setting aside such details and jumping back out to a macro policy standpoint, the reality is that the field of AI research is likely to segregate into those interacting with, tweaking, and combining various forms of API interfaces from those models available from cloud vendors or edge device ecosystems, to those that are working at the most fundamental level where probabilistic algorithm components are becoming more significant to the field, and those in the later camp are going to need to work with and understand quantum computers in a significant way. We will need to find a way to better refine and articulate which of these conventions are more suitable for public research distribution of the kind that has for some time now been common to the field. It may be that broadly public academic works of the kind commonly shared to the arXiv internet portal may discontinue being universally followed by the field. That type of question is an extremely important one, possibly even more so than where research grants are being distributed.

This is not directly responsive to the topic of the RFI, but I would also like to highlight the significance of the energy efficiency of quantum computing towards the challenges of public infrastructure investment needed to support pending further buildout of data center infrastructure leveraging the legacy GPU cluster stack. I have immediate concern that our regional institutions are being asked to make investment decisions for things like power generation infrastructure requiring 30+ year depreciation cycles to build out hardware systems that will reach obsolescence in time scales closer to ~5 years. Traditionally these issues have resolved themselves by GPU vendors rolling out new

generations of hardware that could be substituted for the old, but in an era where we are reaching the limits of Moore's law, we can not take for granted that future generations of GPU hardware will follow similar improvements. The reality is that quantum computers have electricity demand characteristics that are practically trivial in comparison to classical hardware which on current trajectyory has been estimated could increase our national energy consumption by whole percentage points. Obviously quantum computers have tradeoffs (that may vary between vendors) such as noise profiles or exposure to environmental influence that will need to be taken into account for how we allocate and prioritize big compute operations. The reason I bring this up is that as quantum computing is right now becoming capable of outperforming classical conventions of data centers, we need right now to make the decisions on whether we are going to allow our country to be bankrupted by building out huge scales of data centers supported by even huger scales of power generation plants that will handicap our ability to adapt in only a few short years to the certain emergence of quantum computing alternatives, or are we going to take the leap right now and start prioritizing quantum computing as a solution to those operations requiring scaled compute. Even if various domain specific software adaptions aren't fully ready today, they will be soon.

As a separate matter, I also wanted to highlight the work of a Cal-Tech affiliated researcher I recently saw present at the Climate Change and AI workshop at the ICLR 2025 conference related to modern paradigms of AI where the training of large models can almost be taken for granted as solved or at least solvable with sufficient compute. In this setting the bottleneck of model performance may no longer be on how we train our models, it may instead be associated with how do we prioritize acquiring new data streams need to extend our models into new domains. This segment of research also deserves additional attention.

My personal work and research has largely omitted the field of reinforcement learning. I think this modality will be of equal importance in the future. I am not sure how we will be able to "package" RL in a manner resembling the mainstream cloud providers and their LLM commercial offerings. It is an unresolved question and I am interested to see if that may ever become something suitable for commercial settings.