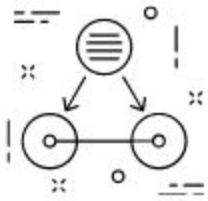


PROTOFECT

Re: **RFI on Developing a Federal AI Standards Engagement Plan**

Protofect Response
Building a rating system for software products using AI

To citizens, leaders, scientists and whom it may concern:

The National Institute of Standards and Technology (NIST) has requested information regarding creating a plan for U.S. Federal engagement in the development of technical standards that will support reliable, robust, and trustworthy Artificial Intelligence (AI) systems and related technologies.

I am Dr. Suman Deb Roy, the Founder and CEO of Protofect LLC, a small data science firm based in New York City. The expertise of Protofect lies in maximizing the performance of data science and AI technologies by systematically designing its interacting modules, predicting adversarial situations, mapping data lineage, optimizing model tuning and prioritizing instrumentation and model explainability.

Protofect is currently building a rating system for software products and data-driven systems that employs artificial intelligence techniques under the hood. Our mission in rating such software products that use AI is primarily to inform consumers, businesses and governments regarding liabilities involved in using the product.

According to PWC, AI is expected to be $15.7 trillion industry by 2030. As the AI revolution booms, potential risks in comprehension, safety, malfunction and misalignment will evolve just as fast as the technology. Unintentional failures will spark the need for oversight — in creation, deployment and proper functioning of AI systems. In the future, before AI products launch in mainstream markets - they will need to be safety rated just as our food products and securities are rated today.

My intention in this letter is to communicate two aspects of the story. First, *the multitude of layered issues that need to be tackled to measure an "AI system"*. Second, *the need and plan of constructing the rating system for AI products*.

The issues surrounding robust, trustworthy and reliable AI systems

The problem of understanding AI reliability, failsafes, robustness and trustworthiness must be considered at not just a holistic level, but at the separate key levels of abstraction of the very ecosystem in which such products flourish. The four levels that immediately come to mind are these:

(1) Computational, (2) Product, (3) Organization and (4) Human impact

This means, AI software is first and foremost computational. There is data science and model embedded in it. But such software, beyond the mathematics, has a design and a timeline, together with deployment mechanism, maintenance costs and an interaction module with the human operator - encompassing the product part. Further, these products flourish in organizations, and are supposed to operate in a certain industry or sector. Finally, the product - through direct or indirect means - has an impact as it touches human life.

We will now attempt to discern the specific challenges in “measurement” of safety in these layers and the different conditions that needs examination.

Computational

1. ***AI Code is not like traditional software code****:* In AI software, a trinitarian concept of the model, training plus new data, and glue-code that stitches the first two together play a vital ever-changing dance. This means unit testing AI modules, unlike traditional software, is quite challenging - because the output function does not maintain linear dependency with the code (like traditional software), but deeply contingent on the model and the new data. Machine Learning (ML) unit tests might pass in one era, but fail in the next. We have no known computational standard of unit testing AI code as of today.

2. ***Verification tools****:* We have a severe lack of verification tools for AI modules. The challenge here is simply related to the nature of “rules of the system”. We assume that software operates on rules and verification verifies if the rules are obeyed. In the good old days of traditional software, these rules were hard-coded. In ML, these rules are learned from the data. ML software builds and rebuilds rules from incomding data. Since there is constant incoming data in most online learning systes, the rebuilding happens in real-time (or in batched intervals).

   These rules comprise the algorithm. Now imagine that algorithms represents the “reality” of what happens inside the “black” box. Every time these rules are rebuilt, the reality alters slightly as new data is encountered. My long term claim has been that ML systems are not *just* black boxes. In fact, the black box nature is possibly irrelevant. In reality (pun intended), ML systems are Schrödinger’s boxes. How do you verify the rules when the rules are, to some extent, fluid? Predictability is a criteria for trustworthiness.

3. ***AI Development Frameworks***: In a similar pattern to software development frameworks (e.g. agile, waterfall), there is an urgent need for ideating what AI development frameworks will look like. Traditional software development frameworks can be unsuited for AI development, due to the 2 reasons previously mentioned. Thus, while we know the general pipeline or life cycle of AI systems (e.g., data ingestions, cleaning, analysis, modeling, deployment) - we must still spend sufficient time in figuring out the best practices around these phases (ideally in a measurable fashion) as the different layers interact, and how to make debugging easier and transparent.

   Similarly, security is paramount. When an AI system is valuable and has significant impact on any aspect of living, it will attract players that do not want to use it in an ethical good-natured way. Thus, just like we have security frameworks, we must spend time in understanding security of AI pipelines, especially upstream - things such as model poisoning or bad reinforcement data. We should model and measure what the collateral impact of failure could be.

4. ***Redefining Accuracy***: Traditionally, academic research has portrayed accuracy in terms of the test and training data. But as we have discussed in the workshops, AI's impact sphere is much bigger than that. Our standards for accuracy measurement of a model are extremely limited to academic purposes, whereas most AI systems, when operating at scale are deployed outside academia. For example, our use of Normalized Discounted Cumulative Gain [1] for search ranking does not capture concepts such as google bombing [2]. Similarly, our accuracy systems for machine translation such as BLEU [3] - do not capture if there is a man-in-the-middle attack during translation. We should revisit and redesign accuracy by measuring more aspects of the ecosystem, and possible threats vectors depending on the sector.

Product

1. ***Integrated and Continuous Data Science:*** Aviation, for all intents and purposes, has become extremely safe compared to the 1950s. While this is definitely due to the engines, pilots, materials etc., the real reward goes to our understanding of fluid dynamics. Thus, the theory of flight was greatly solidified by not only what's visible (the plane), but what's invisible - and surrounds the visible, i.e. the air.

   Data is the air that surrounds visible AI software product. Just like without having a solid understanding of airflow, a plane's flight is unpredictable - similarly, without understanding the flow of data in an ML ecosystem - the system's performance will always remain a mystery. Deployed AI systems, left alone in the wild, have a tendency to deviate from the original goals as it sees new data. A continuous data science effort must completely surround the product to monitor performance, analyze data residues and build theories of behavior. It would also help the situation because often people blame the model for what is just bad data surrounding it.

2. ***Pre-planning for Accidents and Failsafes:*** The big red switch that resets a stage of the AI pipeline, or all of it, is mighty important in certain sectors. Deviation from intended goal cannot be an after-thought anymore. AI developers must spend more time in architecting systems that can be respawned after a failure.

3. ***Certifications***: Builders of specific segments of the AI product would benefit from being certified in their skill to develop - given all the different aspects and complexities of AI systems, just as network engineers or pilots are certified today. Certification agencies could be a wonderful way to attract consolidated guidelines in building future AI.

Organization, Industry and Sector

1. ***Punting aspirational recommendations:*** Conversations around AI safety, security and proper function often get deviated into worries about AGI or suggestions that are directly opposed to fundamental driving forces of ecosystems, free market and western philosophy. For example, a frequent recommendation is injecting a human in "every" loop to verify every stage of building the AI.

   Firstly, while this is a utopian scenario, it could drastically slow down the development of a product, considering companies are under strict timelines to deploy such systems. Secondly, such recommendations, while valuable are applicable with different magnitudes and intensities in unique sectors and thus, must not be generalized. Furthermore, not every recommendation is promising in this current stage as we are too early in the measurement of AI and the parameters of verification.

2. ***Definitions:*** It would be very useful to have a clear definition of issues such as control vs. oversight vs. regulations in relation to artificial intelligence systems. This can reduce confusion in media, and help academia, government and industry collaborate more effectively. A standards body would be paramount in guiding the development of such nomenclature, suited to the sector.

3. ***Not conceding AI leadership.*** A clear characteristic of AI is that it has an exponential effect on whatever it touches. We must be cognizant of the speed at which industries and sector can be reshaped by employing such systems. Intense focus on controlling the development of AI in certain sectors might have the adverse effect of us losing leadership in the field, because we lose talent and opportunity, forced to follow.

   While it can be hard to measure these industry-wide phenomenon, we must research to analyze the speed at which other countries are building AI ecosystems - and lose no ground in chasing aspirational things for the time being. Of course, what looks aspirational at the moment can be practical in the following years due to improvement in measurement science - and it is at that time we must strive to include them comprehensively.

Impact on Humanity:

1. ***Intents and Value systems***: Of all the discussions around AI alignment, perhaps none is more mathematically fuzzy than that of AI with values or ethical AI. Firstly, we must decide on a concrete set of universal values that each sector agrees to adhere too. Secondly, we cannot at present "encode ethics into mathematical models", wherein lies the current disconnect between technical and non-technical folks. We should list the technical limitations that prune ethical aspirations in AI software, and then tackle them one after the other. Perhaps, we could work towards an ethical sci-kit learn type of open source project.

2. ***The Human Condition****:* Advancement of ML and AI should be extended to agencies that surround the human experience, especially religion and law and society. Historians can tell us the cascading effects of new human inventions. Legal folks can tell us the risks of unintentional harm these products could cause. Economists can tell us about the job displacement through AI incursions. Evangelical Christians just released a document comprised of sophisticated points that describes what it means to be human in the age of AI - the ramifications for bias, the workplace, sex and God [4].

   We further need people from different fields (not just computer science) to take an interest in designing the human experience brought upon by AI. This could further alleviate the issue of eroding AI talent.

3. ***AI Task Force****:* Protofect recommends formation of a cross-disciplinary group on the lines of Internet Engineering Task Force (IETF) as an open standards body. AITF would develop and promote voluntary AI standards, and works towards specific protocols in each layer of the ecosystem described above. It is important to note that the IETF itself came to commission as an activity started by the US Federal Government.

Proposal: A rating system for artificial intelligence products

Ratings as an outcome of measurement science. From if-then-else loops to deep neuro-evolution [5], AI technology will arrive in different forms. The measurement of various components in an AI system requires careful sensing, estimation and judgement of how the modules act and interact. An example of different modules Protofect inspects and measure in the AI pipeline is shown in the Addendum.

Protofect's rating system is algorithmic and for products that employ AI or predictive/prescriptive data modules in any capacity for its functioning. It will help institutions to become more transparent to consumers, investors and oversight committees when using AI, by providing clarity about components of AI software, evolution from traditional technology and its direct or indirect impact on digital ecosystems.

We rate AI products algorithmically based on numerous factors such as data sources or ingestion stability to learning and model accuracy, using multiple proprietary tools such as deep questionnaires, adversarial datasets and interrogative APIs. The engine is mathematical, not based just on anecdotal evidence or media coverage. We also consider the idea behind the AI product, its history, the technology, the implementation, the company's maturity and responsibility via compliance, employees, the market and critical infrastructure and impact.

As a start, we have begun measuring some of these attributes of the AI ecosystem along the following dimensions:

Dimension	Properties	Description
1. Training Data	Ownership, Noise and Velocity	Training data is what a model is built upon. Multiple factors that go into understanding the assumptions in collection, storage, cleanup, class weights, pre-processing privacy and lineage of data.
2. Learning	Offline, Online and Reinforced	The learning algorithm entails not only the model represents the data distribution, but also its tuning mechanism, update sequences, and maintenance from feature erosion.
3. Algorithm Complexity	Linear, Non-linear, Evolving	Every model lives on an axis from simple to complex. Models could be linear, non-linear or reinforced over time. The more complex they are, the harder it is for humans to comprehend the underlying engine and prediction trail.
4. Benchmarks	Settings, Performance, Optimization	When benchmark datasets are available, we will test the performance of this system under varying levels of difficulty. Benchmarks create consistency, expected behavior and reveal known failures and risks.
5. Real World Test	Beyond Simulation and Emulations	While simulations and emulations can provide powerful results, running on real-world, live data for a period of time can test an algorithms robustness to multiple factors that cannot be imagined up in simulations.

6. Debuggability	Diagnosing and Surgical Tuning	Learning algorithms can be hard to debug because prediction often depends on the data, and not on the software (which could be probabilistic). Data residues left in the prediction pipeline reveal signals to diagnose and surgically improve the model.
7. Turing Strong	Performance against Humans	One of the major controversies surrounding AI is that it displaces human jobs. This efficacy claim is to be judged by a consumer. Will the human serve better? Or the AI? And what happens to the displaced human? Certain market leaders might need to spend time on this dimension and model the displacement vs. improvement.
8. Bias Control	Liberty and Equality	Many AI algorithms are accused of biased predictions. There is a general understanding that these might be introduced by the designers, or by the training data. However, a bias meter is essential so we can understand and rectify the deviation from intended behavior.
9. Democratic API	APIs as Freedom of Information about Model	While not everyone wants to release information the training and prediction pipelines, APIs that can interrogate the model should be available - as a way for bidirectional accountability
10. Company Responsibilities	Maturity and Duties	While you do not expect startups to have entire research departments dedicated to certain ethical AI missions, a mature public company, especially ones in regulated industry, should comply with these. The older the company, the greater its funding, the bigger its market share - the more is its responsibilities.
11. Audits	Trusting the Trusted	Because AI code and design is extremely valuable property, it will be treated as a trade secret. But a team of

		auditors should have the ability to go through the prediction pipeline, so its responsibilities are met to the consumers and society. Just like the food industry and securities were eventually asked to follow a set of guidelines for the overall benefit of society, so should many civilization critical technologies be responsible to their actions.
12. Market Impact	Mission Criticality	Mission critical AI has a greater responsibility than its peers, as its decisions can have deep impact on human life, values and ethics.

Protofect's rating system, once built, will discern the relative risk that the AI product will fail to align with the original intention of the creators, operators and beneficiaries in that market. It addresses the possibility that the AI software's obligation to the business, users, markets and society will not be honored. Our ratings reflect both the likelihood of failure and the corresponding loss suffered in defaulting.

We thank NIST for hosting the AI standards workshop, and taking the lead in getting the community together for one of the most important discussions of our times. We would also like to invite NIST members and interested parties in the community to suggest, propose, assist and use the rating architecture as a tool for AI standard.

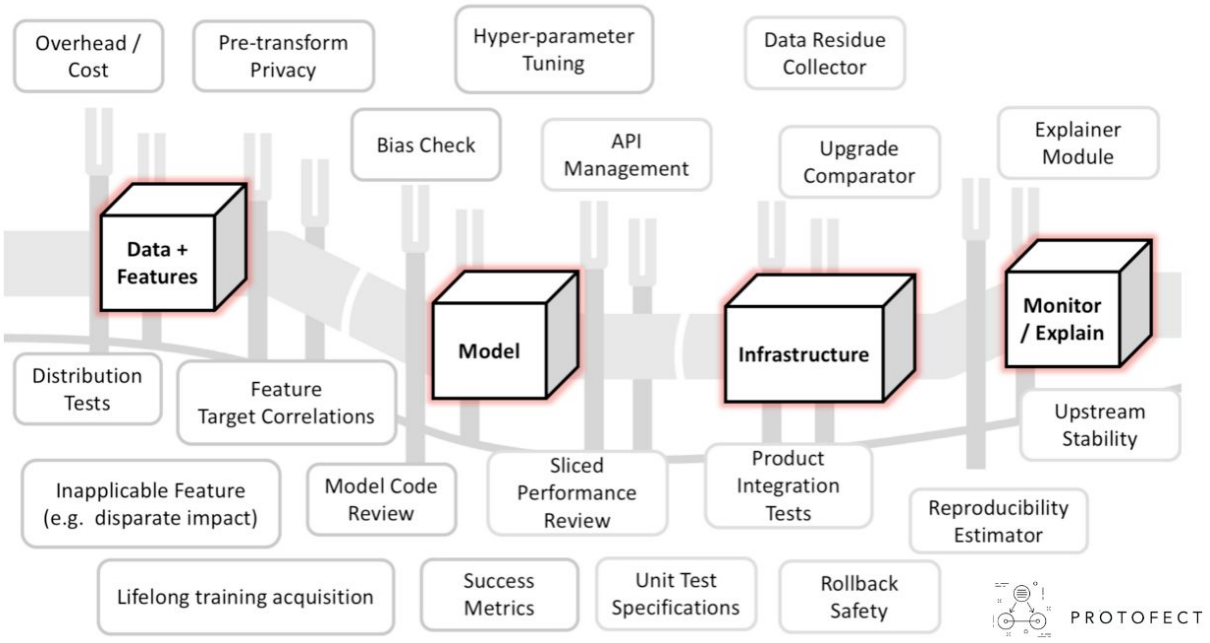
For questions/comments and getting in touch, please email roy@protofect.com

---

References:

[1] http://proceedings.mlr.press/v30/Wang13.pdf
[2] Bar-Ilan, J. (2007). Manipulating search engine algorithms: the case of Google. *Journal of Information, Communication and Ethics in Society*, *5*(2/3), 155-166
[3] https://www.aclweb.org/anthology/P02-1040.pdf
[4] https://erlc.com/resource-library/statements/artificial-intelligence-an-evangelical-statement-of-principles
[5] Tirumala, S. S., Ali, S., & Ramesh, C. P. (2016, August). Evolving deep neural networks: A new prospect. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 69-74). IEEE.

Addendum:


Overhead / Cost
Pre-transform Privacy
Hyper-parameter Tuning
Data Residue Collector
Bias Check
API Management
Upgrade Comparator
Explainer Module
Data + Features
Model
Infrastructure
Monitor / Explain
Distribution Tests
Feature Target Correlations
Upstream Stability
Inapplicable Feature (e.g. disparate impact)
Model Code Review
Sliced Performance Review
Product Integration Tests
Reproducibility Estimator
Lifelong training acquisition
Success Metrics
Unit Test Specifications
Rollback Safety
PROTOFECT