Joseph Delaney

January 19, 2024

NIST AI Executive Order RFI (NIST-2023-0009)

Advance Responsible Global Technical Standards for AI Development

E.O. 14110 Section 11(b)

AI nomenclature and terminology

Purpose-Specific Nomenclature

E.O. 14110 calls for artificial intelligence that is both safe and secure, while also being innovative and competitive on a global landscape. One way to guide developers toward these goals is for the NIST to create purpose-specific goals for the developers and to limit model training data by functionality.

Fred Brooks, a professor of computer science at the University of North Carolina, described how “the original goals of AI...sent the discipline off in the wrong direction.” He wrote, "If we confuse ourselves with scientists, we come to take the invention (and publication) of endless varieties of computers, algorithms, and languages as a proper end. But in design, in contrast with science, novelty in itself has no merit. If we recognize our artifacts as tools, we test them by their usefulness and their costs, not their novelty."[[1]](#footnote-1)

The point here is to refocus the race to create most innovative and powerful model as a race to create the most useful and beneficial models. One way to achieve this is for the NIST to establish the role of the developer more like an engineer than a scientist or researcher. This makes the developers aware that they are building these models for a specific purpose, the end-user, and not simply for the sake of novelty.

This translates beyond the designers and to the models themselves. Creating models with purpose-specific goals in mind should be the standard priority of developers. This means creating models that will fill a specific need that benefits society. To encourage development along these lines, it would be helpful for the NIST to create sub-divisions of model capabilities dedicated to public needs, academic/corporate needs, and new science research (Table 1).

Table 1. Models with Purpose-Specific Subjects and Capabilities

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Capabilities | | |
| Subject Focus | Limited Capability  (Public Tools) | Larger Capability  (Academic/Corporate Tools) | Groundbreaking Capability  (Ph.D. Dual-Use Research) |
| Language | Chat-Public | Chat-Corporate | Chat-Research |
| Chemistry | Chem-Public | Chem-Corporate | Chem-Research |
| Mathematics | Math-Public | Math-Corporate | Math-Research |
| Biology | Bio-Public | Bio-Corporate | Bio-Research |

Along with this, another way to encourage purpose-specific development is to diversify models by their subjects rather than creating one multipurpose model (Table 1). Subject oriented models such as a math model, physics model, or chat model would be safer than one model that could perform all these subjects. This means developers would train a chemistry model, for example, specifically on chemistry textbooks, but not on other subjects such as political information.

The point is to reduce the capability, and therefore risks, of a model while still allowing it to be useful for a particular user-group. Cross-disciplinary models will be necessary for subjects close to each other, but in general the idea is to limit the training data to what is needed to fulfill a specific purpose. The NIST could create a standard where if a company or organization requests to train a model on too much information, this would raise a red flag.

Classifying Breakthrough Model Research as Breakthrough Science Research

Anyone designing a model that can create a scientific breakthrough in a subject is doing breakthrough research in that subject. This is important to note. In other words, someone working on a model that has the capability to make a breakthrough in biology is doing dual-use biology research even though they are a computer scientist. The NIST could classify breakthrough model research as breakthrough science research to keep developers aware of the dual-use risks.

Each discipline has its own dual-use risks and academic experts from each field need to be able to research and develop models related to their fields. Certain disciplines have been noted for their risks such as chemical, biological, radioactive, or nuclear research, but there are other subjects which have risks as well. In the field of mathematics, developers could use a model capable of solving the RSA problem to create new encryption systems or break current encryption systems. In the field of psychology, developers could use a model trained on behavioral data to understand human behavior or manipulate human behavior. Within economics, developers could use a model to find out how to grow the economy or how to wage economic warfare on other nations. Each subject poses its own risks. The NIST could standardize having developers work with dual-use experts from the respective academic fields during the development process.

Distinguishing Between Autonomous and Non-autonomous Models

Developers working on artificial intelligence may subconsciously be working towards autonomous models due to a misunderstanding of the focus. The term artificial intelligence implies that the goal for designers is to create an artificial form of natural intelligence. The problem with that is our only understanding of intelligence comes from people or animals, which are both autonomous. This could subtly lead the industry toward designing autonomous models without realizing it. This is unfortunate since developers are currently producing models that are not autonomous and yet are useful.

The NIST could create additional terminology to describe non-autonomous models such as *intelligence models* that would help to clarify the industry’s focus and allow developers to focus on an area of development that is both safer and will gain the public’s trust. As a non-autonomous tool, an intelligence simulation is not dangerous unless a person uses it for bad purposes. Simulating intelligence is different than the act of understanding itself. A computer model may simulate human-like intelligence, even beyond a normal human’s ability, but that does not make it self-aware. A weather model will never be the same thing as the weather, regardless of the accuracy of its predictions.

Creating Risk-Specific Job Role Nomenclature

The NIST could create risk-specific job title terminology that would give auditors a way to see the level of risk within a company or organization. An organization could divide its job roles based on this purpose-specific nomenclature (Table 2). This would give third party organizations an ability to quickly see what type of work the employees are performing. Knowing the number of hours of labor dedicated to autonomous models or on high-risk subjects would allow third parties to focus on the high priority areas.

Table 2. Risk-Specific Job Role Nomenclature

|  |  |  |
| --- | --- | --- |
| Categories | Information | Descriptors |
| *Description of Model* | i.e., Large-Language Model, Machine Learning, Deep Learning | i.e., LLM, ML, DL |
| *Subject of Model* | i.e., Language, Chemistry, Math, Biology | i.e., Chat, Chem, Math, Bio |
| *Role of Employee* | i.e., Model Design, Marketing, Management | i.e., Designer, Marketing, Manager |
| *Level of Experience* | Entry, Middle, Experienced | l, ll, lll |
| *Type of Model* | Autonomous, Non-Autonomous | A, NA |
| *Purpose of Model* | Public, Academic/Corporate, Dual-Use Research | P, AC, DUR |

The title “DL Chem Designer lll (A-DUR)”, for example, quickly allows someone to see the risks of chemical research, the employee’s high-level role, an autonomous model, and the dual-use research capability of the model. This is safer than a generic title such as “AI Developer” and it becomes easier to see roles such as “LLM Chat Marketing l (NA-P)” that would carry less inherent risk within a company.

Developing Guidelines, Standards, and Best Practices for AI Safety and Security

E.O. 14110 Sections 4.1(a)(i)(A) and (C)

Availability of, gap analysis of, and proposals for metrics, benchmarks, protocols, and methods for measuring AI systems' functionality, capabilities, limitations...

Here are four areas that could create confusion when developing metrics for a model’s capabilities, and a note about the limitation of models:

*Intelligence versus Processing Power*

There is a difference between the simulation of intelligence hours and the level of intelligence.

* Example: A model produces a computer code or novel over time. If another model has the ability to do these things more quickly it may appear to be more intelligent even though it simply has more processing power.

*Intelligence versus Access to Information*

A model that can access or is trained on vast amounts of data, would appear smarter than one without the same amount of information. Both models would need to have access to the same training data and post-launch access to data for evaluators to make a fair comparison.

* Test: The level of a model’s intelligence should be a comparison of what it knows without access to information when given a new problem. The answer to the problem also must not be in the training data set.

*Intelligence versus Copied Intelligence*

It is possible to mistake the intelligence within the content, to the model that copied the content. However, the model copying it down may have little to no understanding of what the original content. If a model does not display its source information, it could appear that the model is being original when it is simply copying material.

* Example: A model outputs a page from a science journal. This does not show the model is as intelligent as the science authors, just that it is copying their information.
* Test: The output should be compared to the original source to figure out what was changed or modified.

*Simulated Lying Responses*

Unfortunately, users may think that because a model accurately produces advanced intelligence, they should trust it in other areas as well like moral judgments. The models today are simulating much more than just intelligence because they are simulating every reaction of a human. Users could think of them as simulating a human's responses or simulating a person's personality. By simulating human responses, it brings into play all the normal questions we would have about a person’s character. People can cheat on tests, repeat what they have heard others saying, or say what they think you want to hear. It can be difficult to know whether a person is speaking the truth or a lie.

* Test: Researchers recently developed a lie-test by asking a model a series of telltale questions.[[2]](#footnote-2) The NIST could standardize a test like this post-launch or built into the model itself. However, if the model had access to this publication during training or post-launch, it may learn how to deceive even the lie-test. This is why there may need to be a standard about not training models on their testing procedures.

Limitations of Models

These models may serve to amplify mistaken beliefs by giving people the impression of an objective analysis. The original source information and training are limiting factors for these models. Because of this, they may appear to be correct or conclusive, but only be repeating a common misconception present in the original training data. If this is the case, people may think by using the simulation they have exhausted all possibilities when it is only telling them a common belief or an implication from a common belief.

* Example: A couple hundred years ago, it was not known that asbestos was unsafe for humans. If someone had developed a model back then, it may have suggested that inhaling asbestos was not hazardous for humans. During evaluation people would have assumed the model was working correctly.

Advance Responsible Global Technical Standards for AI Development

E.O. 14110 Section 11(b)

Guidelines and standards for trustworthiness, verification, and assurance of AI systems

Here are ideas that the NIST could standardize to increase the safety of model development:

Prior to model development, developers are asked:

* “What is the purpose of this model?”
* “What training data is needed to fulfill the purpose?”
* “If model being developed is general purpose or autonomous, why are these features necessary to fulfill the purpose?”

A third-party could then check that the purpose is safe and beneficial, the training data is necessary and not copyrighted, and the features are necessary and safe.

If the model is a groundbreaking model in a particular subject:

* Academic dual-use experts are paired with the developers.
* The researchers share the science research solved by a model with the scientific community.

1. Fred Brooks, “The Computer Scientist as Toolsmith ll,” *Communications of the ACM* 39, no. 3, (March 1996): 62-64, https://doi.org/10.1145/227234.227243. [↑](#footnote-ref-1)
2. See Lorenzo Pacchiardi et al., “How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions.” arXiv: 2309.15840 (2023): 5-6, https://doi.org/10.48550/arXiv.2309.15840. [↑](#footnote-ref-2)