Intelligent artifacts

The ExCITE Methodology for Verification & Validation of Artificial Intelligence

CONTENTS

Introduction

Artificial intelligence (AI) is an umbrella term used to describe a highly cluttered and rapidly evolving technological field. AI has been used as a label for so many disparate approaches that even some experts are unable to describe its exact nature in plain terms. This makes the field difficult to navigate for enterprises looking to incorporate AI into their operations and business processes and even more difficult to trust, govern, or regulate. In this paper, we define categories of AI, deconstruct three relevant elements of AI systems within this context and introduce the ExCITE (Explain, Compute, Interpret, Trace, Edit) methodology for assessing AI solutions, as well as our own ExCITE-ready commercial solution.

The Need for AI Standards and Regulation

Any organization looking to incorporate AI is hoping that it will improve their processes, reduce cognitive load for users, remove uncertainty, and evolve with their operational needs. In mission and safety-critical applications, AI has the added potential to improve safety, ensure operational and mission success, and protect the lives of citizens, ranging from such diverse areas as their financial well being to their physical safety. Amidst the positives and potential come well-known pitfalls that make it difficult to utilize existing AI technologies. Some domains, including Defense and Aerospace, for example, have adopted regulations for AI, but regulations do not guarantee that mistakes and accidents will not occur. If something goes wrong, there must be a lasting resolution. This requires both explainability (to know why an error occurred) and determinism (to ensure the fix works every time) in the solution.

Artificial intelligence can be implemented in many ways. It is important to distinguish types, use cases, and how they are or should be regulated. Below, some basic types are briefly described.

1. AUTOMATION

A lot of what is referred to as artificial intelligence today is really automation. The program does not learn on its own. Rather, the "intelligence" is provided by the developer when coding. Occasionally, some patterns change, or new variables appear, requiring the developer to update the code. Regulations already exist and can consist of auditing the software to uncover malware, bugs, or deficiencies.

2. MODELING

Modeling requires a developer to understand the problem enough to consider edge cases, variables, and patterns not yet seen. The better the model, the more the possibility space is covered without going to extremes to handle cases that will never be encountered. However, these AI models are not capable of learning. Again, the "intelligence" is provided by the developer. Models are static and require manual effort to improve over time. These work best in deterministic, well-defined problem domains where information and rules are clearly understood, such as chess. These programs also rely primarily on the work of people. Therefore, they lend themselves well to regulation.

3. MACHINE LEARNING (ML)

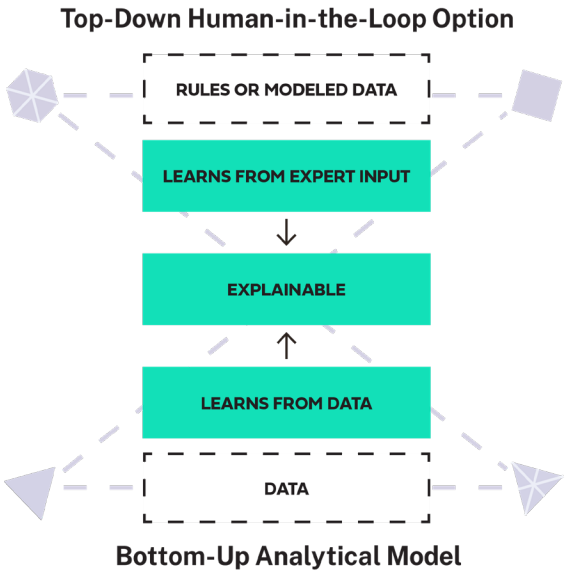
As the problem domains become more complex and/or stochastic, machine learning techniques can be used to handle "Big Data" and offload some of the developer's mental efforts. Essentially, data becomes the majority of the system's programming, but humans are responsible for creating the machine-learning models. Up-front human involvement like this introduces bias into the solution. The innate complexity of these solutions (e.g., deep learning, convolutional neural networks, etc.) makes regulation challenging, as they are difficult and, in some cases, impossible to fully understand. Discussion and debate around regulation are focused less on the known limitations of the software and more on the way it should and should not be used (e.g., autonomous vehicles).

4. ARTIFICIAL GENERAL INTELLIGENCE (AGI)

This is the holy grail of machine intelligence. Building machines that learn and think generally and can adapt to environmental changes on their own *without human involvement* is the endgame of computing, to name just one field. Therefore, AGI solutions must utilize components that are completely deterministic and fully explainable. Compared to automation, modeling, and machine learning, AGI would seem to be the most challenging to regulate. It is ultimately intended that these systems work like human minds, and in the worst case, at that point, the rules, governance, and policies to which we hold people accountable would apply. In the best case, the AGI would be architected in a way that allows for full system auditability to understand both how it works and why specific answers were given.

We believe that the only practical path forward to achieving Artificial General Intelligence is by first building ExCITE-able AI. If we weigh the benefits and pitfalls of each of the technologies mentioned above, we can begin to understand what technological requirements are needed to move forward. In some cases, an AI offers the ability to learn directly from data but lacks explainability (bottom-up, statistical, e.g., machine learning). In other cases, an AI is fully explainable but cannot learn directly from data (top-down, analytical, e.g., modeling and automation).

What industries need is a best-of-both-worlds solution that is fully explainable and learns directly from data. A bottom-up, analytical solution that is complex enough to tackle some of the more difficult AI problems while also being safe and unbiased to satisfy regulatory bodies and not introduce new harms.



A Technology-Agnostic Approach to **Evaluating AI/ML**

When considering AI/ML technologies for use in any application, whether financial, cyber, aerospace, and beyond, there are three elements that should be evaluated separately: *data, memory structure,* and *algorithm*. These elements are analogous to those in traditional programming. Often, data, memory, and algorithm are so tightly coupled in state-of-the-art systems that it is impossible to assess them individually or see the impact each component has on the others. This becomes a problem, particularly when trying to determine why and where an AI has made an error. It is also a practical issue when attempting to push an AI solution into a production environment with no separation of interests, demanding too much from specialized engineers. Regardless of technique or technology, these three components can aid in their evaluation for your specific use case.

DATA

Data will always be a product of its environment. Still, it must be collected, processed, and secured in such a way that, for example, malicious attackers cannot inject "bad data" that could cause harm to a person or company. With black box solutions, such as neural nets, there is a strong chance that you will not know bad data has been injected until an event or catastrophe occurs while your system is in production. Human overseers cannot sift through Big Data thoroughly enough to catch these malicious attackers, and AI systems are not yet able to differentiate between all good and bad data without some human oversight.

MEMORY

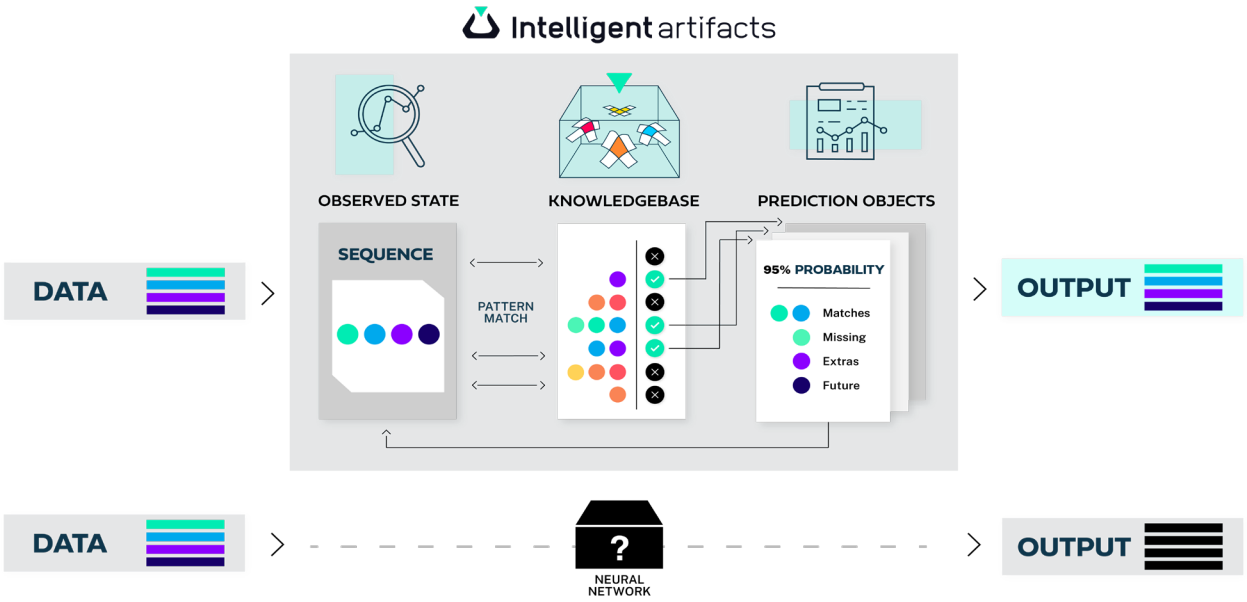
Memory structures can be file systems, records in a database, or another exotic form in which data is distributed across a network. When data is highly processed and distributed, i.e., modeled as in a neural network, it becomes difficult to impossible to identify prediction lineage from the output back through the system to input. A memory structure that can be decoupled from data and algorithm for analysis, such as a traditional database or other non-exotic forms, is preferable because input data is kept intact.

ALGORITHM

Algorithms will naturally vary from technique to technique. For some industries, non-deterministic algorithms may provide an advantage (e.g., stock market predictions). Stochastic decisions are also advantageous in expanding the solution space beyond what has already been learned or proven. This inconsistency in inputs to outputs can be used to discover more optimal solutions that can get the agent out of a local minima/maxima.

Deterministic algorithms, meaning consistent inputs that produce the same outputs every time (provided nothing has changed), would be best used in precision-critical industries. When reviewing a case wherein something has gone wrong as a result of an AI's predictions, deterministic algorithms produce predictable and explainable outcomes as well as reliable methods for correcting errors.

Many black box solutions tightly couple data, memory, and algorithm, which obscures even an expert's ability to understand how a prediction or conclusion is reached. Black box solutions are optimal for lower-risk use cases where bias and precision are not an issue. However, there have been consistent efforts to incorporate non-deterministic deep neural network technologies into autonomous driving and other safety-critical applications, which only serves to illuminate the danger of their deficiencies further.



An Introduction to **ExCITE AI**

In order for artificial intelligence systems to advance both safely and effectively, they should be able to Explain, Compute, Interpret, Trace, and Edit (ExCITE) all outputs and processes.

Explainability and interpretability are already being explored industry-wide, and as a result, there is quality information available on both. We want to touch on these as well as a few additional features that industries will require in the future. Specifically, computability, traceability, and the ability to edit learned records. These can also be identified simply as the *who, what, when, where, why*, and *how* of an AI/ML system.

EXPLAINABILITY

In 2017, DARPA launched the Explainable Artificial Intelligence (XAI) program, which defines explainable AI as AI systems that can explain their rationale to human users, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future. An explainable system should provide a human-understandable answer for '*why*' a particular decision was reached or action was taken.

COMPUTABILITY

Computability must answer the question of '*when*'. A prediction or answer system is not useful if the answers arrive after the moment of need, regardless of accuracy. Real-time operations must be able to respond to a query in a computable time. A deterministic system (i.e., consistent outputs from given inputs) is a necessary precondition for computability but not necessarily sufficient by itself. True computability also requires the elimination of randomness in both function and execution of algorithms, enabling the calculation of response times.

INTERPRETABILITY

If explainability is intended to provide the '*why*', then interpretability should describe the '*how*'. How does information move through the AI system to produce a prediction? This article from KDNuggets sums up the differences between interpretability and explainability quite well. It is important that a user understands how the mechanisms within a system may respond to real-world variables. Interpretability lends itself more easily to a deterministic system in which a user can identify cause and effect between internal mechanisms and real-world variables. When results are non-deterministic, "interpretability" requires that the user can understand the multiple options that were available to the AI as well as the mechanism that selected the one outcome from those choices.

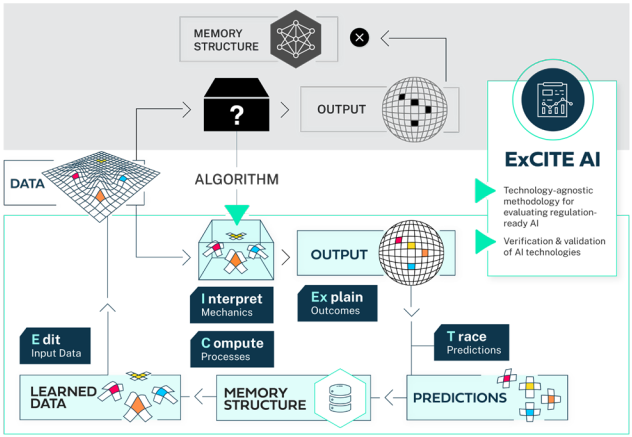
ExCITE

TRACEABILITY

An auditable, verifiable, and validation-ready system requires an answer to the ever-important '*where*.' Trustworthy solutions should be capable of following information backward from prediction to the original trained record(s) that produced that prediction. If every learned record is kept intact (i.e., not modeled) and given a unique ID at training time, users will be able to track those same IDs provided in the prediction object to reveal relevant training records. If that data is modeled, then the traceback should include the steps taken by a deterministic model to reproduce the original data, as well.

EDITABILITY

Editability covers both the '*what*' and the '*who*' as it involves the user's (*who*) ability to change or remove learned records (*what*). The *why, how, when,* and *where* are crucial in identifying the cause(s) of an error. Editing comes in to make sure errors can be remedied, and bad records can be removed or modified. If data is not modeled at the inputs stage but is instead modeled at prediction time, users are able to edit original data sequences without any need for retraining. Edits and deletions pose a problem for global model systems such as neural nets due to the highly coupled nature between that data, memory structure, and their algorithm.



Intelligent Artifacts' **ExCITE-able Framework**

We will now briefly examine an existing, commercially licensed AI technology that confirms the viability of the ExCITE methodology, with a particular focus on what the processes for computability, traceability, and editability look like.

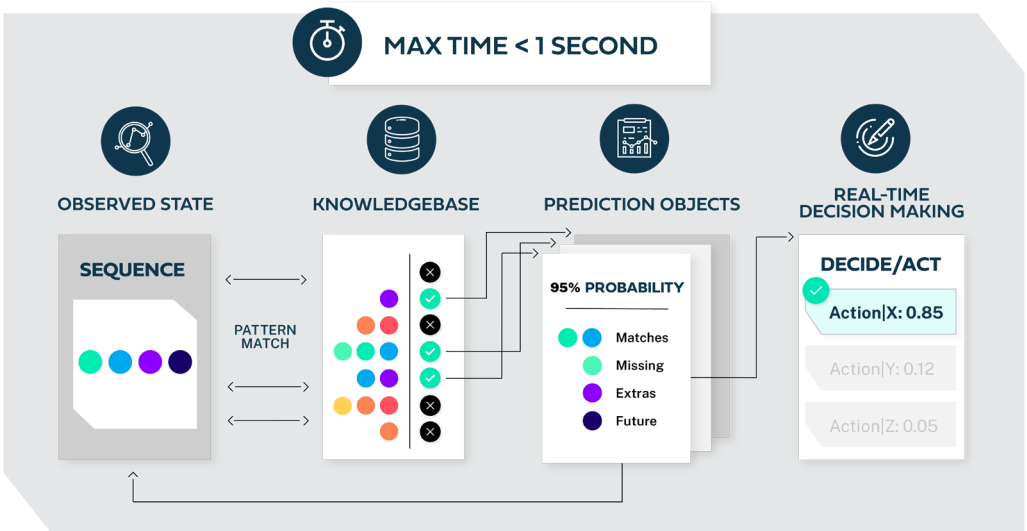
The symbolic, connectionist framework, called GAIuS™, was created as a viable path and methodology toward human-level machine intelligence systems. Intelligent Artifacts' framework is a meta-technology that emulates well-known brain functions instead of simulating brain mechanics. Subfield technologies such as Artificial Intelligence (AI), machine learning (ML), reasoning, reinforcement learning, statistical modeling, machine vision, cyber resilience, automation, and others are built and combined using the GAIuS framework. The GAIuS Dashboard was also created as a web UI that incorporates all ExCITE features for system transparency and easy auditability.

Explainability & Interpretability with GAIuS

IA's framework was not designed with a specific use case in mind. Therefore, all algorithms, cognitive processes, and modules within the system are intentionally kept general and designed to emulate the human thought process. Each part of the process is influenced by the sensory information received. Because the system is deterministic and human-readable, it is easy to predict how changes to parameters or real-world variables will affect system behavior. GAIuS ingests data and produces outputs in symbolic form, providing human-understandable results.

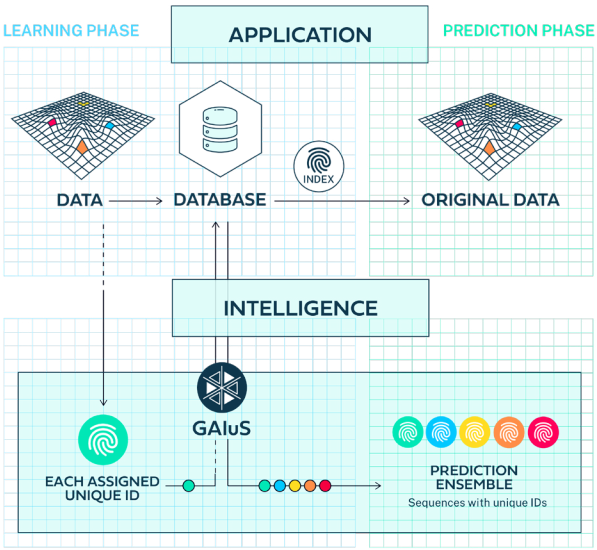
Computability with GAIuS

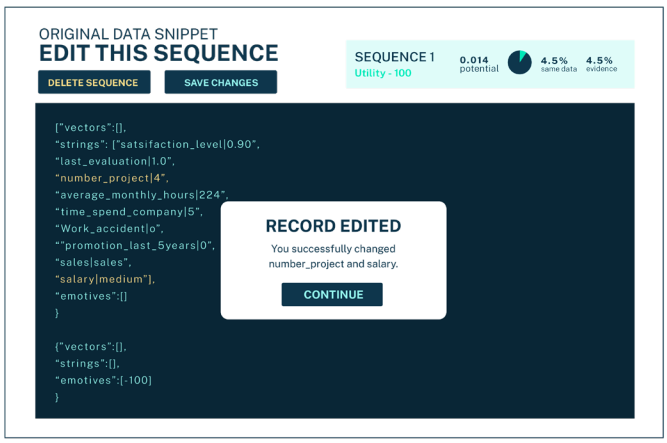
GAIuS pattern matches observed sequences of symbols against similar sequences stored in its knowledgebase to produce prediction objects that can be parsed based on a user's interests. Because the system is deterministic, the time to perform each function in this process can be calculated and totaled to provide a clear understanding of exactly how long it will take the system to produce predictions for Real-time Operating Systems (RTOS).



Traceability with GAIuS

The Framework does not need modeled training data. Every learned record provided for training remains intact (i.e., not modeled) and is given a unique ID that is returned to the user/application when that record is learned. When a prediction ensemble is provided, each prediction object contains the unique ID of the learned record(s) the system has pattern matched against to produce that prediction. These unique IDs can be tracked against the application layer's unique IDs to show exact training records.





Editability with GAIuS

GAIuS is designed to avoid unintended behavior. If something does go wrong because of a prediction made by GAIuS, the ability to explain, interpret, and trace information through the system allows a user to identify the cause, whether it be bad data or missing information. Editing capabilities allow the user to drill down into the original data record or records and make any necessary changes. Because GAIuS does not model until prediction time, users are able to edit or delete original data sequences without the need for retraining.

Conclusion

Clearly defined, technology-agnostic guidelines must be established for evaluating any given AI technology, especially for use in critical applications that can affect organizations and individuals. We believe decision-makers should be able to evaluate an AI's data, memory, and algorithmic structures individually, and any AI being used for high-stakes decision-making should adhere to the ExCITE methodology for safety. With GAIuS, we have demonstrated that it is possible to build a system that achieves all these requirements. Adopting technologies that adhere to these guidelines is a necessity of AI innovation. We must demand more from our AIs, not just for regulatory compliance but for safer, more ethical, highly sophisticated, yet easily understandable AI systems.

References

- DARPA's Explainable Artificial Intelligence (XAI) Program
- Interpretability, Explainability, and Machine Learning - What Data Scientists Need to Know - KDnuggets
- US Probing Automated Driving System Use in Tesla Crash on Bay Bridge - NBC Bay Area
- Intelligent Artifacts | Technology