Verification & Validation of Project Management AI[1]

Bob Prieto

Chairman & CEO
Strategic Program Management LLC

Previously I have written about the potential use of AI in the management of large complex projects[2] and several issues which arise in such usage and outlined the need for effective verification and validation (V&V). In this paper some areas of special concern are highlighted and thoughts on how to approach aspects of the V&V process offered. This paper considers the work of others in the broader verification and validation community as well as my own thoughts as to V&V in AI enabled project management. The intent of this paper is to foster a discussion of this important area as various project management AI efforts move apace.

Defining V&V

There exists a need to continuously define verification and validation since a range of thoughts exist across the literature. This definitional challenge is addressed by Gonzalez (2000)[3] and can be seen in the differences between the Institute of Electrical and Electronic Engineers (IEEE) and the US Department of Defense (DoD).

Verification deals with satisfying specifications. Verification involves the structural correctness of the knowledge base, that is to say it is internally consistent and complete. Among the challenges are a taxonomy which addresses key milestones and activities in the project execution process. Such a taxonomy must be capable of transcending project types and a tendency to redefine it for each project or data subset must be avoided.

Verification of AI enabled project management systems must move beyond static rule testing given the non-deterministic nature of AI programs. We must check for inconsistencies and incompleteness; discrepancies, ambiguities and redundancy. Verification must ensure that all portions of the rules base are exercised in testing. Testing with known results may leave us susceptible to errors because of weak or incomplete coverage of the test set. Verification of AI enabled project management systems requires

[1] How to cite this paper: Prieto, B. (2019). Verification & Validation of Project Management AI. *PM World Journal*, Vol. VIII, Issue X, November
[2] Prieto, B. (2019). Impacts of Artificial Intelligence on Management of Large Complex Projects. *PM World Journal*, Vol. VIII, Issue V, June.
[3] Gonzalez, A., Barr, V. (2000). Validation and verification of intelligent systems – what are they and how are they different? Journal of Experimental and Theoretical Artificial Intelligence, Volume 12

us to think about underrepresented data subsets and late stage or other temporal failure regimes. For example, how does our AI model behave when confronted with a test data set of all project successes and asked to look for failure?

Accuracy depends on training data set and data inputs. Issues of correctness, completeness and appropriateness of source data quality can be a failure point. There is a need for automated data quality checks.

Project management AI systems will require coverage measures to assess the quality of the domain populations (training and test) and meta-knowledge to provide guidance on:

- fitness for a specific use case
- likelihood they are in the training population
- how representative the test set is of the intended population

There is a need to develop standardized measures and ontologies to allow the proposed project management AI systems to be evaluated.

Verification must happen first, and only then validation.

Validation involves exercising the system, testing the project management AI. This is a dynamic process in which we test for functional correctness, comparing the behavior and predictions of the project management AI to the real world, or at least our interpretation of it. We define failure. We define what an acceptable level of predictive confidence is. We define what constitutes a valuable lead time over timely human prediction. All these

> **Verification** – ensuring intelligent system conforms to specification; its knowledge base is consistent and complete within itself
>
> **Validation** – process of ensuring that the output of the intelligent system is equivalent to those of human experts given the same inputs

considerations and other go into our validation of the developed AI. Data validation and model selection have not gotten enough rigorous attention to date and much more remains to be done.

One other consideration deals with validation for intended use. This is discussed later.

Role of Project Management AI

Project management AI is founded on predictive analytics[4]. It reflects uncertainty, best described as a confidence level. Confidence will change over time as the project progresses and new data is considered by the AI model.

Types of analytical models

Descriptive analytics – historical data analysis and visualization

Predictive analytics – predicting future based on past data

Prescriptive analytics – prescribing course of action from past data

Project management AI is at its best diagnostic, but with an expert viewpoint. It is a decision aid, part of the broader project management system[5]. It helps the project team make better decisions

It is unlikely that users will trust an AI system no matter how impressive some of its demonstrations of competence may be without some causal explanation of its behavior.

Verification and validation is part of that causal explanation process.

Trust = fitness for use, reliability and robustness

Role of V&V

Verification and validation is about establishing trust. Users must trust adaptive, non-deterministic or complex AI systems.

Building this trust is made harder recognizing that AI does not lend itself to checking using conventional V&V methodologies. AI is complex and therefore harder to understand and test. Methods developed for assessing normal software still apply but are best described as good but not sufficient.

Traditional testing is typically scenario based, seeking to consider the maximum number of potential situations with the minimum number of tests. AI testing is compounded by its

[4] Iyengar, V., Subramanian G, S. (2018), The Right Testing Strategy for AI Systems, An Infosys Viewpoint
[5] Prieto, R. (2019). Proper Reliance on Artificial Intelligence in Project Management; *PM World Journal*, Vol. VIII, Issue VIII, September

complexity and could conceivably take more resources than its initial development. In AI the number of potential situations is too large

AI V&V will increasingly rely on model checking. The distinct advantage of model-based AI systems is that the high level description is the system. Declarative knowledge describes facts and relationships within a domain making it easier to understand/use/communicate, and importantly build trust. Simple queries for this declarative model reduces V&V efforts, but only if we have necessary confidence in the model and the data it was trained on.

Several early efforts in project predictive analytics have shown an ability to detect failure but fallen short on predicting success. Is the inability to predict success a model weakness? Does an opportunity exist to test the sensitivity to failure by changing one variable at a time over a range? What does model performance and our V&V activities tell us about our own project execution methodologies?

Clustering analysis will allow us to find related parameters, yielding new insights for the project manager. AI enabled planning and scheduling tools can be treated similar to model checking in that they all explore a state space described by a model.

V&V of AI enabled project management systems should ensure that we see sanity properties such as consistency, absence of ambiguity or expected properties such as functional dependency between variables. The resultant tools must provide trustful diagnosis such as being able to infer accurate and sufficient information on the state of project from its observed behavior.

Bias

In his paper[6] Ghoneim addresses the subject of bias and the notion of artifacts or artificial patterns that are caused by deficiencies in the data collection process. For example, in our quest to understand project failure we preferentially populate the training data set with failed projects, not providing the tool with a representative sampling of successful projects. Alternately, the data quality of many of our worst performing projects excludes them from our training data, limiting our ability to potentially see the worst of the worst.

Other examples of potential data deficiencies could include other representation (or over representation) of one client or client type, geographic region, or project delivery method (E,P,C; EPCM; EPC; PPP/PFI)

In general we can define several types of bias potentially impacting project management AI. These include:

[6] Ghoneim, S., 5 Types of bias & how to eliminate them in your machine learning project; Data Science in the Real World

- **Sample bias** – Sample bias arises when our training and test data do not represent the project environment in which we are deploying the project management AI tool. For example, we train the tool on oil and gas projects but then seek to deploy it on infrastructure projects. This suggest a potential need to parse our universe of data into robust subsets trained for the intended use. While a broadly trained tool may provide different and even potentially more valuable insights than one trained on a subset of the data, it would not be unreasonable to expect the more specialized data subset to provide higher confidence insights. An evaluation of these two approaches is warranted as the industry moves forward. We must extend validation theory to define data validation – does the data provide a valid representation of the problem space?

- **Exclusion bias** – During initial tool development and proof of concept testing it would not be unreasonable to limit the number of project features/ fields of data to be considered. As tool development proceeds it will be more valuable to include more fields of data for consideration. For example, when considering overall success/failure of large complex projects one may note that delayed completion of the process engineering stage correlates very strongly with overall project failure. As we expand the range of data considered it would be advantageous to incorporate more data related to the precursors of process engineering, external influencers, and actual process engineering performance. Such inclusion would allow for diagnosis of process engineering challenges earlier in the execution process, focusing corrective action at an earlier point, even before overall project diagnosis might otherwise have flagged a problem. In the data collection and preparation stage we must take care in 'cleaning' any data and be aware that early reporting period data may actually reflect plan versus actual as the project management systems are being stood up, or worse, the project is experiencing poor project startup.

- **Observer bias** – The training and testing teams see what they expect to see. Their thinking is anchored and weaker but important insights are overlooked. In the training and testing stage it is important to screen for potential biases by involved humans. Later in tool deployment we see a different form of observer bias, denial, which acts to reject what doesn't fit our current mental models.

- **Prejudice bias** – This arises in our selection of training and testing data where we unconsciously reflect stereotypes in our data. There is a need for a more even handed distribution of examples. For example, not enough successful projects or not enough in a given size range that is within the scope of intended

use. Proper representation at the extremes of the intended use range is particularly important.

- **Measurement bias** – This reflects a systemic bias in how/what we measure. For example, we measure funds expended versus commitments made; actual productivity versus rate of productivity improvement; RFIs or number of holds versus drawings issued for construction.

Verification and validation of project management employed AI needs to include an important element of bias testing. Model predictions must be consistent for all possible inputs. Did the system learn something we are unaware of[7] such as failure linkage to size which becomes problematic in a different currency or after escalation? We need to test for true positives, true negatives, false positives and false negatives.

Special Challenges of Adaptive Models

AI systems are prone to error due to their complexity. AI, specifically AI based on machine language, is non-deterministic, or said differently there is uncertainty on the system's future behavior. It can behave differently for different runs.

Non-deterministic choices come from incoming external events, scheduling of concurrent tasks, and intentional random choices. This results in exponentially many possible executions, creating a state space explosion.

Non-determinism can be described as either:

- **External**, resulting from inputs from the system's environment. From a modeling standpoint these can be assumed to be controllable, although they are less so controllable in the world of real project execution. Their ability to be controlled from a modeling perspective should allow us to test the impact of externalities in advance? Much of the AI project management effort to date has relied on internal project data but earlier insights into actual project performance may be gained by considering the impact of externalities.

- **Internal,** arising from our project execution system and methodology itself. Concurrency (concurrent execution) of project activities or binary or stochastic (making random or pseudo-random choices) limits our ability to test such systems.

Non determinism is a V&V issue for AI

[7] Kohli, P., Dvijotham, K., Uesato, J., Gowal, S. (2019) Identifying and eliminating bugs in learned predictive models, Deep Mind blog

Concurrency is a natural source of non-determinism since the order and timings of independent processes can vary. We see this in project execution and recognize that a new set of risks emerge as we increase project concurrency through techniques such as modularization and parallel execution. It is important that we understand and minimize concurrency risks.

Stochastic non-determinism can arise in AI enabled project management systems through adaptation of system behavior as a result of new project data. We can learn from the changes these data additions have on our model, understanding how the system responds to these new internal choices allowing us to increasingly turn them into external choices and opportunities. Within a non-deterministic program a small number of assumptions control which non-deterministic option the program will take. We see this in master-variable scheduling and machine learning in other domains. This provides a project management opportunity to focus on control of a few variables[8].

Adaptive systems change in response to expanded data sets[9]. In a sense they can improve themselves, becoming more 'confident' and executing faster. This evolving nature creates a challenge for verification and validation. Adaptation may render obsolete any pre-adaptation certification, perhaps necessitating dynamic comparison between the current model being used and the certified model. Understanding how the model migrates over time can provide new insights.

Test Methods

We have previously discussed the need for quality, unbiased data. But now we need enough of it to adequately test the model and its adaptation with new project examples. In a sense the first step in the validation process is ensuring the adequacy of the training and test set. The model developed from the training data will be initially tested with data not used to build it. We must remain cognizant that a learned model in one domain may not apply to another.

Further insight, especially with respect to homogeneous subsets, can be gained using N-way cross validation. Under this approach training data can be sorted into N buckets. For each N, the other N-1 buckets have a learned algorithm that is then tested on the N^{th} bucket. The prediction of error rate is the average of the classification accuracy and provides a descriptive parameter on model performance.

[8] This is referred to as the funnel assumption. It is argued that searches through a space containing uncertainties, most of the reachable conclusions will be reached via a small number of "master variables" in a "narrow funnel".
[9] When the data mined indicates systemic changes, the self-modifying prognostic system refines the previously developed algorithm; Pullum, L., Darrah, M., Skias, S., Tso, K. S., Tai, A. T. (2004) Developing a Data Driven Prognostic System with Limited System Information, IEEE

Additional insights can be gained when assessing different adaptation mechanisms (M). Repeating the N-ways cross validation M times results in NxM training and test sets. The resultant mean and standard deviation compare with a *t* test with N degrees of freedom. Smaller samples exhibit fatter tails with the distribution approaching a normal distribution as sample size grows.

Heterogeneous data sources can be a failure point. There is a need to check the model's ability to handle heterogeneous data during comparison

How does learnt theory[10] change over time as more data is processed? When does learning stabilize (more data does not change the model)? Have we collected too much data in a given domain? Sequence studies can aid in answering these questions as well as consideration of artificial distributions such as the planned behavior.

Project management AI will benefit from an anomaly detector that tests if new data is different than data that has been previously managed. A pre-filter could reject new input if too anomalous while a post-filter could recognize unusual output and stop it from effecting the rest of system. This addresses the need to ensure stability of the output theory.

Such an anomaly detector could also be relevant for an ongoing fit-for-purpose test, letting us know if the current model deployment is out of bounds of fit-for-purpose.

Testing methodologies should investigate the value of sensitivity analysis, robustness to small perturbations in inputs, and identify useful metrics for AI based software and validate their value as predictors of important cost, performance and reliability measures. Alternative models, such as regression models[11], may be derived from sufficiently large data sets and can be used in conjunction with the AI developed model and provide a reasonableness test.

Summary

Proper reliance on artificial intelligence in project management requires strong AI predictive tools, with known confidence levels at various time frames (less confident prediction of failure early on but with a strengthening predictive confidence as more time lapses), including:

- Transparent and robust AI algorithms, trained on known, relevant data sets and validated for intended use.

[10] Menzies, T., Pecheur, C. (2004) Verification and Validation and Artificial Intelligence
[11] Rushby, J. (1988) Quality Measures and Assurance for AI Software; Technical Report CSL-88-7R performed for the National Aeronautics and Space Administration under contract NAS1 17067

- Knowledgeable deployment of validated AI to use cases verified to be consistent with the validated AI.
- Recognition of AI limitations due to excluded data (external ecosystem data) and an assessment of the relevance of its consideration in the particular use case (project)

Verification must ensure the developed intelligent system conforms to its specification and that its knowledge base is consistent and complete within itself. Special attention on data is required including understanding and confirming any of a range of potential biases. In addition, transparency and verification of user data rights emerges as a core issue.

Individual users who contribute data to a multi-enterprise training data set must retain sufficient rights over their data while the broader (multi-enterprise) insights gained are derivative. User data must be protected and users must maintain control over access and usage of their data.

Other data, of uncertain or unknown provenance should not be used in an AI algorithm or service without verifying rights, applicability and appropriateness for intended use, and embedded bias in data.

Users should ensure that both their data and any AI algorithms they make available are robust against manipulation.

Validation is the process of ensuring that the output of the intelligent system is equivalent to those of human experts given the same inputs. The robust validation regime that is suggested is necessary to achieve a high degree of confidence in the validity of the AI algorithm and therefore confidence that it has been represented and used with a high degree of explainability. Explainability should have as a minimum the following attributes:

- Decision making process should be explainable
- Recommendations should include sufficient explanations, data used and limitations, reasoning
- AI decision processes should be verifiable
- AI intent should be transparent
- AI algorithms may be powerful and scalable but also transparent to inspection
- Understand how added data changes expected outcomes

Project management AI holds great promise. It demands that we treat and use these tools in new ways, with new training and new mindsets.

About the Author

Bob Prieto

Chairman & CEO
Strategic Program Management LLC
Jupiter, Florida, USA

Bob Prieto is a senior executive effective in shaping and executing business strategy and a recognized leader within the infrastructure, engineering and construction industries. Currently Bob heads his own management consulting practice, Strategic Program Management LLC. He previously served as a senior vice president of Fluor, one of the largest engineering and construction companies in the world. He focuses on the development and delivery of large, complex projects worldwide and consults with owners across all market sectors in the development of programmatic delivery strategies. He is author of nine books including "Strategic Program Management", "The Giga Factor: Program Management in the Engineering and Construction Industry", "Application of Life Cycle Analysis in the Capital Assets Industry", "Capital Efficiency: Pull All the Levers" and, most recently, "Theory of Management of Large Complex Projects" published by the Construction Management Association of America (CMAA) as well as over 600 other papers and presentations.

Bob is an Independent Member of the Shareholder Committee of Mott MacDonald. He is a member of the ASCE Industry Leaders Council, National Academy of Construction, a Fellow of the Construction Management Association of America and member of several university departmental and campus advisory boards. Bob served until 2006 as a U.S. presidential appointee to the Asia Pacific Economic Cooperation (APEC) Business Advisory Council (ABAC), working with U.S. and Asia-Pacific business leaders to shape the framework for trade and economic growth. He had previously served as both as Chairman of the Engineering and Construction Governors of the World Economic Forum and co-chair of the infrastructure task force formed after September 11th by the New York City Chamber of Commerce. Previously, he served as Chairman at Parsons Brinckerhoff (PB) and a non-executive director of Cardn0 (ASX)

Bob can be contacted at rpstrategic@comcast.net.