Patterns

CellPress
OPEN ACCESS

Opinion

Validating and Verifying AI Systems

David J. Hand[1] and Shakeel Khan[2,*]
[1]Department of Mathematics, Imperial College, London, UK
[2]Chief Data Officer's team, Her Majesty's Revenue and Customs, London, UK
*Correspondence: shakeel.khan@hmrc.gov.uk
https://doi.org/10.1016/j.patter.2020.100037

AI systems will only fulfill their promise for society if they can be relied upon. This means that the role and task of the system must be properly formulated; that the system must be bug free, be based on properly representative data, and can cope with anomalies and data quality issues; and that its output is sufficiently accurate for the task.

Introduction

AI systems are becoming ubiquitous in modern life, ranging over medical diagnostic systems, financial trading algorithms, driverless cars, customer engagement systems, and countless other areas. Sometimes performance of the AI is critical, in the sense that a patient could die, an accident could occur, or a business collapse if incorrect decisions are made. So a key question is: can you trust your AI? How do you know it is doing what you want it to do?

Such high level questions have several aspects:

(1) Has the objective been properly formulated?
(2) Is the AI system free of software bugs?
(3) Is the AI system based on properly representative data?
(4) Can the AI system cope with anomalies and inevitable data glitches?
(5) Is the AI system sufficiently accurate?

Positive answers to all of these questions are needed to have full confidence and trust in the performance of the system. The aim of this paper is to look at these questions from a high level perspective. A more technical discussion is given in Menzies and Pecheur.[1]

AI is not unique in having to address these questions, and analogous questions are central to many other domains. This has inevitably meant that various terms are used for the different aspects. In health and psychology, for example, the term "reliability" is used to describe the reproducibility of a measurement result under different conditions (so it refers especially to questions 3 and 4), while the term "validity" is used to describe whether the measurement procedure is tapping into the right concept (question 1). In software testing, a distinction is made between "validation" and "verification." Validation refers to checking that the system specifications satisfy the customer's need (question 1), while verification is checking that the software meets the specifications (especially questions 2, 3, and 4). Informally, validation is sorting out that you are answering the right question, and verification is ensuring that you find the right answer to that question. In machine learning, validation is often used in the narrow sense of ensuring that the predictions are sufficiently accurate (think of the phrase "cross-validation," for example). This is the subject of question 5 and might better be called "evaluation." It has various aspects, briefly described below.

In this paper, we will use the term validation as a shorthand term to cover all aspects.

Question 1 involves mapping real-world questions (with all the ambiguity, uncertainty, complexity, and wooliness typical of the real world) to a formal mathematical description, which can be described in a programming language. This is basically an assessment of conceptual accuracy, asking whether the AI system is addressing the right problem. Answering question 1 may not involve data at all but could simply require elaborate exploration of design documents and specifications in an effort to detect problems, anomalies, or oversights, as well as the possibility of a system being presented with unexpected conditions. The complexities of the real world mean that guaranteeing the adequacy of this mapping will often be impossible, and the best one can do is to try to think of all possible scenarios that could arise. Question 1 also involves ethical issues, for example the question of whether or not an AI personnel selection system discriminates. This example also illustrates the complexity of the challenge, since there are several, mutually incompatible definitions of discrimination, one of which must be chosen for the system.

Questions 2–5 involve more mathematical exercises. Given a (hopefully) well-defined problem from question 1, the aim is to explore whether the solution (the AI system) answers it. In extreme cases, automatic theorem proving systems can be used, but the apparent finality of formal mathematical proofs should not seduce one into a belief that the system is necessarily doing a good job: the importance of positive answers to *all* of the questions is illustrated by the cautionary comment from Xie et al. to the effect that "formal proofs of an algorithm's optimal quality do not guarantee that an application implements or uses the algorithm correctly, and thus software testing is necessary."[2]

Validating AI algorithms is tougher than validating conventional algorithms because the former may have the capacity to adapt. Indeed, that is often the essence of such programs and is what provides the "intelligent" in "artificial intelligence" and the "learning" in "machine learning." Sometimes this is described as meaning their behavior is "non-deterministic," because it depends on external events or other changing circumstances (not least, random internal aspects, as with simulated annealing,

Image 1: Presentation by Professor David Hand at the Validate AI Conference, November 2019

genetic algorithms, and stochastic approximation). The adaptability of such systems and their flexibility in responding to external events can lead to a state space explosion.

Aspects of Validation

Question 1 is clearly very context dependent, and there is little we can say in general about that—at least in a short space. The other questions, however, hinge on two main aspects: the data and the algorithm. We shall look at these aspects separately.

Data quality is a perennial issue for all quantitative disciplines, in particular including statistics, data mining, machine learning, and artificial intelligence (see, for example, Breck et al.[3]). As various aphorisms (e.g., "garbage in, garbage out") attest, it is a truism that the validity of the results of an analysis depends on the quality of the input data. And while it might be the case that a system works perfectly with perfect data, it is a brave assumption to suppose that the data the system encounters in practical application will always be perfect. And note that very large or rapidly streaming datasets cannot be checked by hand.

In general, while algorithms might be robust to some data quality issues, there will be others to which they are highly sensitive, and there will be breakdown points in terms of the extent of poor quality that can be handled. In some cases, such breakdown points involve only a tiny percentage of the data. For example, Karmon et al. change just 2% of the pixels in an image, none of them over the main object, and almost always fool image-recognition systems.[4] It is also important to note that data might be perfectly fine for some purpose but poor for another—validation must be applied with the right objective in mind.

Data often arise from multiple sources, linked, merged, or otherwise combined by the AI, and the different datasets might have different degrees of quality—and of compatibility. Validation exercise should explore these aspects.

In addition to obvious data quality issues, there are challenges of non-stationary problems—so-called population drift—where the nature of the underlying population changes over time. For example, as more driverless vehicles appear on the roads, so the average expected behavior of vehicles will change. Validation should consider the complete life cycle of the system. In general, the data to which a system is exposed during a validation exercise should span the entire space of scenarios the system is likely to encounter, insofar as this is possible. Statisticians have always cautioned about extrapolating beyond the data, but the autonomy of an AI system means it might well encounter novel situations.

Validation of the algorithms themselves means confirming that they solve the problem presented to them—really questions 2–5. Apart from formal mathematical proofs, which can be used in limited circumstances, the most common strategy is to embed the AI in an artificial environment that generates simulated data of the kind it is likely to meet. Clearly the efficacy of this depends on how well the simulated data reflect real data, complete with anomalous data points, etc. As we noted above, it is important to generate extreme cases and span the space of types of situations. This can be a challenge for AI systems because of the diversity of different scenarios they might encounter. The struggles to develop fully autonomous driverless vehicles have illustrated this. But even using real test data can run into problems. Real data have often undergone some prior selection procedure such that they may not properly represent the population that the AI will be dealing with. For example, in retail credit, the training data will typically be past customers, but they will have been enrolled as customers through some selection process and are unlikely to be representative of *all* future applicants.

There is an analogy to stress testing that plays a major role in validating more general systems, such as financial models used by banks, though there the main aim is to look at response to extreme conditions. As the Bank of England puts it: "Banking stress tests assess how banks can cope with severe economic scenarios. We look at banks' resilience, making sure they have enough capital to withstand extreme shocks and are able to support the economy."[5] Sensitivity analysis is another related idea.

In many situations, validation involves running the algorithm on cases where the "right" answer is known, to see whether, while the question might be properly formulated and the AI system might function as intended, it might simply not be very good. For example, even though a medical condition has been properly described and a system built to diagnose on the basis of appropriate symptoms and test results, and even though there are no errors in the programming (e.g., it uses a well-established and bug-free logistic regression algorithm), a medical diagnostic system might

Image 2: Validate AI Conference program brochures, November 2019

misclassify a large number of cases simply because the implied decision surface is too rigidly constrained. This sort of problem has been the focus of a vast amount of work with a wide variety of performance criteria being used (see, e.g., Hand[6]). This abundance can conceal the importance of ensuring that an appropriate criterion is used. For example, by far the great majority of assessments of diagnostic performance of machine learning algorithms use misclassification rate, even though this is often inappropriate (since it treats the different kinds of misclassification as equally serious). There is often also a trade-off between the robustness of a system, meaning it does not react wildly to slight changes of the data, and accuracy, meaning that it does not adapt sufficiently to changes in the data.

A common strategy used in validating (or, perhaps more appropriately, "evaluating") such algorithms is a cross-validation or holdout strategy, in which available data are split into two sets: one to construct the algorithm (e.g., estimate parameters) and the other to test it. Note, however, that this assumes stationarity of the underlying populations. In many real situations, future data are unlikely to be drawn from exactly the same distribution as the design data, so a misleading impression can be gained.

The point about including extreme cases in the validation exercise raises the question of whether an AI knows its limits. If sufficiently anomalous data arise, the system should recognize this and stop, rather than simply continuing regardless. Think of an autonomous lawn mower stopping at the edge of the lawn.

Conclusions

Considerable effort has been made in recent years to enable AI systems to "explain" their decisions. This is partly driven by legal requirements. For example, the European Union's General Data Protection Regulation (clause 71) says "[automatic processing of data] should be subject to suitable safeguards, which should include … the right to … obtain an explanation of the decision reached …" (though this is controversial—see Wachter et al.[7]). Beyond any legal requirements, however, explainability is often associated with superior generalizability and greater robustness because of the natural regularization implicit in human understanding and mental modeling of phenomena. Moreover, while it is true that opacity of a system means it is difficult to tell how and why it is going wrong when it errs, it is also true that explainability and the way it is implemented will depend on who the explanation is for.

AI systems typically work in a social context. So validation needs to do more than examine systems in isolation. It is also necessary to explore how people work with and react to such systems and, indeed, to see how other AIs work with each other. The risks are illustrated by the behavior of correlated financial trading systems: if one says "sell," other similar systems are likely to do so as well, possibly leading to a financial crash.

Further discussion is given in the white paper produced after the Validate AI Conference. The conference and subsequent related activities have been organized to mobilize academic and practitioner groups to advance academic research and applied methodologies in AI systems of validation. Further information can be found at https://www.validateaiconference.com.

REFERENCES

1. Menzies, T., and Pecheur, C. (2005). Verification and Validation and Artificial Intelligence. Adv. Comput. *65*, 153–201.
2. Xie, X., Ho, J.W.K., Murphy, C., Kaiser, G., Xu, B., and Chen, T.Y. (2011). Testing and validating machine learning classifiers by metamorphic testing. J. Syst. Softw. *84*, 544–558.
3. Breck, E., Polyzotis, N., Roy, S., Whang, S.E., and Zinkevish, M. (2019). Data validation for machine learning. Proceedings of the 2nd SysML Conference. https://mlsys.org/Conferences/2019/doc/2019/167.pdf.
4. Karmon, D., Zoran, D., and Goldberg, Y. (2018). LaVAN: Localized and Visible Adversarial Noise. arXiv, 1801.02608 https://arxiv.org/abs/1801.02608.
5. BoE (2020). https://www.bankofengland.co.uk/stress-testing.
6. Hand, D.J. (2012). Assessing the performance of classification methods. Int. Stat. Rev. *80*, 400–414.
7. Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. Int. Data Privacy Law. https://doi.org/10.2139/ssrn.2903469.

About the Authors

David J. Hand is emeritus professor of mathematics and senior research investigator at Imperial College London, where he formerly chaired the Statistics Section. He a past president of the Royal Statistical Society and is a fellow of the British Academy. His books include *Dark Data*, *The Improbability Principle*, *Information Generation*, *Intelligent Data Analysis*, *Artificial Intelligence and Psychiatry*, and *Principles of Data Mining*.

Shakeel Khan is Artificial Intelligence (AI) Capability Building Lead at Her Majesty's Revenue and Customs (HMRC) and is co-founder of the Validate AI Conference initiative. He has been a great advocate of AI deployment in HMRC, wider government, and tax administrations globally. Prior to this he worked in financial services leading major machine learning initiatives. He has worked closely with academics thoughout his career to champion AI innovation.