Evaluating the gap between hype and performance of AI systems

Azamat Kamzin, Prajwal Paudyal, Ayan Banerjee, Sandeep K.S. Gupta
IMPACT Lab, Arizona State University, Tempe, Az
{akamzin, ppaudyal, abanerj3, sandee.gupta}@asu.edu

Abstract

AI systems are generally evaluated based on some performance metrics like accuracy or Root Mean Squared Error (RMSE). While using such metrics are meaningful, they do not fully capture the needed evaluation. As more AI systems are getting deployed in real-world situations, a variety of other AI evaluation criteria must be considered. In this paper, we discuss such situations when traditional techniques of AI evaluation based on static datasets fall short. Specifically, we discuss effects due to transparency of participant bias, disclosure of priorities, choice of metrics that may not be robust over time, the need to explain results of AI and compare explanations and the need to account for effects due to adaptive behaviors of humans.

Introduction

Artificial Intelligence (AI) implies a form of data-driven learning that enables the automation of tasks that require various levels of human cognition. AI components are increasingly being included in applications that are now poised to directly affect human life such as autonomous cars, closed-loop blood glucose control systems, or cardiac diagnostic systems. Such large scale deployments have seen a common artifact: a significant gap between expectations from an AI system and its practical performance. This gap has influenced human behavior as a result novel use cases have been seen on the field for which the AI system was not evaluated. Such use cases have often led to safety hazards. The gap in expectation and practical performance always exists for any automation system regardless of the presence of AI. However, with the incorporation of AI, this gap has broadened due to many reasons some of which are directly related to the evaluation methodology of AI systems and are the main focus of this paper. To highlight these issues, we consider three practically deployed automation systems that have some AI components in them:

i) Automated Sign Language Tutor: Learn2Sign (Paudyal et al. 2019) is a mobile or web-based application that teaches users American Sign Language (ASL). The application shows a video of a chosen ASL sign, the user can review the sign by browsing the video multiple times, the user can then practice the sign execution by using the webcam and recording a video of the execution. An AI engine matches the practice video with a tutorial video and evaluates the execution as correct or incorrect. ii) Automatic control of insulin infusion using artificial pancreas (AP): Using a continuous glucose monitor (CGM) through the wireless interface the controller tracks the tissue glucose levels every 5 mins. The controller then computes the insulin level to be infused to the human body to keep the blood glucose level within a given range. Several types of controllers are used including PID (FDA 2016), model predictive (Brown et al. 2018), and self-adaptive (Messer et al. 2018) controllers that learn from CGM excursion history. The controller is semi-autonomous, where mealtime insulin infusion is manually announced and managed by the user. iii) Model-driven smart cardiac monitoring: GeMREM (Nabar et al. 2011) is a cardiac event monitoring system, that develops an individualized generative model of the human heart by learning the model parameters from historical data. The model is used to check current deviation and if no deviation, then no new data is transmitted and the model itself can be used to regenerate the signal. However, if any cardiac events result in a change in beat shape then a new model is automatically learned to adapt to the new event and create a report of any adversities. Utilizing these examples, we focus on the following reasons for the gap in the hype and performance of AI systems.

- Lack of transparency in the evaluation of the AI system: The overview of evaluation methodology of widely marketed AI systems are publicly available in many different forms through technical publications, safety approval reports, or manufacturer manuals. However, several important details that can potentially explain the gap in reported and practically observed performance are missed. For example, the cohort used for evaluating the performance of Medtronic 670G system reported their CGM data voluntarily and can be expected to be in better adherence with the protocol. However, the users in a free-living scenario may not be expected to follow the protocols strictly. There is no evaluation of the system that checks for nonadherence to different aspects of the protocols and their consequences.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- Marketing of anything AI creates hype which evaluations do not match: To the common user, usage of the word AI is often vague and they can be misled to have higher expectations from the device through improper marketing strategies. In the case of AP systems, the gap between marketed performance and observed performance is so drastic that there are several initiatives such as OpenAPS (https://openaps.org/), where the common user is encouraged to design their own AP. One of the main reasons is although the Medtronic 670G AP promises to have a high percentage of time in normoglycemic range, in evaluation the system was configured to reduce immediate life-threatening effects of hypoglycemia than more commonly occurring hyperglycemia which has long term effects. The system is not evaluated for a user who requires more aggressive insulin delivery to control hyperglycemia.
- Evaluation metrics are not human-centric: In the case of cardiac event detection and monitoring mechanism such as GeMREM, the evaluation typically involves matching the generated signal with raw data or counting the number of critical events detected and matching with manual annotations. Metrics such as mean square error or true positives false positives are typically used. Although these metrics are perfect for evaluating the performance of the system, they are not diagnostically relevant. For example, a 10% error in heart rate measurement is insignificant, however, the same error in Q T interval computation can be crucial in the diagnosis of diseases. For the AP example, percentage time in hypoglycemia (reported in Medtronic 670G) may be a great metric to evaluate the overall performance of the controller, but hypoglycemia event statistics including the number of events and duration of each event is not reported and they may be more relevant for the analysis of serious insulin management issues. Initial parameters can be learned by mining them from output traces generated by human usage, thus enhancing human-centric solutions (Lamrani, Banerjee, and Gupta 2018; Gupta, Banerjee, and Lamrani 2019).
- In the ASL tutor (Paudyal et al. 2019), the evaluation is fundamentally a recognition task, where practice video is compared with a tutor video. As such F1 score, precision, recall can be good metrics to evaluate the recognition capability. However, such numbers have no consequence on the learning outcomes of the students. A recognition result of correct or incorrect execution of a sign should also be accompanied by a description of the reason for the recognition result, and if wrong how to correct the execution.

As such, the gap between the expectation and practical performance of a system should be explicitly tackled by the evaluation system. This is significant for optimal user experience of AI systems, and if not addressed can lead to novel use cases that are may often have unintended harm to the user. For example, in the case of AP devices such as Medtronic 670G, there are reported cases where a user provides phantom carbs (med 2018) to trick the system in providing extra bolus insulin input. In other examples, a misconception regarding autonomy of a car can result in an inattentive operator leading to potentially fatal consequences (ube 2019). In this paper, we consider several examples of AI-enabled systems and evaluate their evaluation to report best practices (BP) for evaluating AI systems that can potentially address and explain the gap between hype and practically observed performance of AI systems.

BP1: Transparency on the effects of evaluation methodologies on participant bias

Participant bias is a common term used in clinical studies for evaluating the effect of an intervention. This generally means that the subjects of the clinical study change their behavior based on their expectations of the outcomes desired by the study co-ordinator (Smith and Noble 2014). Although this effect is not typically considered in the evaluation of AI systems given the increasing trend of human interaction with AI components, the effect of an evaluation methodology on participant bias has to be evaluated.

Participant bias can be of different forms depending on the evaluation methodology. For example, the evaluation of an AI system can require the subjects to follow a protocol as in the case of AP devices. In such a methodology, the participant whose CGM data is monitored can be overly attentive to the details of the protocol. This may result in a use case that is not often observed in practice and devoid of natural variations of human users.

Moreover, for systems such as ASL tutors (Paudyal et al. 2019), the evaluation is based on correct recognition of gesture executions when a participant practices. Here participant bias can act both in favor, the user may be extra attentive to execute a gesture perfectly, or against the performance of the system, the user may intentionally provide wrong executions during the testing phase.

Participant bias can, in fact, be utilized to explain the gap between expected and practically observed performance. This is because a biased participant can potentially provide controlled experiments. The effects of such experiments can be used as templates in a root cause analysis framework to explain the failures of the AI system.

BP2: Disclosure of Priorities of Evaluation Criterias

The first step in evaluating an AI system involves generating use-cases, functional and non-functional requirements and the design of experiments. Non-functional requirements such as safety are typically generated through close collaboration with manufacturers and regulators and often the user is included either through surveys or through consultation with domain experts. However, for a complex AI system, a multitude of safety requirements may be extracted through a hazard analysis step. For agile and cost effective evaluation of the AI system requirements are typically prioritized. The typical priorities include immediate safety hazards that can have fatal consequences. As a result, some of the more long term safety risks may be ignored. For example, incase of the Minimed 670G AP approved by Food and Drug Administration (FDA), the primary safety requirement was the avoidance of hypoglycemia. But recent post-market evaluations (Leelarathna and Thabit 2018) show that postprandial-

hyperglycemia is a significant problem for the approved device which has long term risks of high HbA1C levels (Landgraf 2004) and potential organ failure (Gerich 2013). Post prandial hyperglycemia is not discussed in approval documentation of the Minimed 670G (FDA 2016). A significant side effect of this drawback is that the Medtronic 670G controllers are designed to be conservative in automated insulin delivery to avoid hypoglycemia, while the more aggressive bolus infusions are left for manual interventions from the user. The problem is errors in such interventions can introduce significant risks of hypo or hyperglycemia which the controller may not be evaluated to handle.

Priorities in evaluation criteria are typically not explicitly disclosed to the user and depend on several factors that may be internal to a AI system manufacturing and business unit. Priorities are guided by the cost of evaluation, time taken to perform the evaluation, estimated severity with respect to safety violations, and many more factors. *However, these priorities determine the gap between what the AI system can do and what it is evaluated to do.*

If such priorities are not disclosed, then a human user may end up using an AI system under conditions for which it is not thoroughly evaluated. This may result in unprecedented behavior potentially leading to unexpected poor performance. Hence a fundamental requirement for an evaluation method of AI systems has to be disclosure of priorities for evaluating functional and non-functional requirements.

BP3: Choosing Robust Metrics

AI systems can be evaluated using various metrics like accuracy, mean squared error, precision, and recall. Some of these metrics like the mean squared error or cross-entropy are utilized directly to optimize the machine learning components of many AI systems. AI engineers have realized over the years that different types of mistakes can have different costs, so there are also examples of using cost functions that incorporate these by means of penalizing either the False positives or False negatives more heavily. For example, in the case of cardiac event detection or prediction systems, the mean square error metric specifically is used quite frequently in works that take a signal processing based data-driven learning (Gee et al. 2016). However, clinical research (Zigel, Cohen, and Katz 2000) has shown that mean square error metrics are by no means useful for caregivers or clinical researchers for the diagnosis of diseases.

Other metrics such as computation time are used as 'satisficing' metrics that help engineers select between two otherwise equally performant models. However, these metrics do not incorporate many other aspects of AI systems such as interpretability to the end-users or depletion of performance over time. Various online learning systems can be monitored and constantly reevaluated over time i.e. as the underlying environment of operation changes. Current evaluation techniques do not take into account such effects during the initial evaluation of the systems. This could be remedied by having evaluation systems that run multiple candidate systems in shadow mode as the initially chosen model performs inferences. This will allow a better comparison of how the performance of the various systems depletes over time. The other aspect that will be discussed in Section 3 is the interpretability of the system. While it can be agreed that more interpretability is generally better than less, the amount of trust we can attribute to interpretations of various classes of AI techniques has not been evaluated. For instance, different systems can give varying explanations for the same decision as has been noted by Wojciech et al (Samek 2019). In addition, researchers rarely test AI systems according to how they integrate into the overall decision process. For instance, in the case of a recommendation system that works together with various other AI systems to improve the user experience, the choice of the system itself is made prior to integration without much regard for possible interactions between components.

BP4: Evaluation in Terms of Explainable Concepts

With the rapid advancement in AI and wide range deployment of such systems in human-interfaces applications, there is a growing need for greater transparency and trustworthiness of the systems in the eyes of human-end users. Increasingly advancements in AI mean the growing complexity of underlying algorithms and the black-box nature of these systems, requires to reevaluate existing evaluation approaches that have driven our current progress in AI. There is a need for end-user to be involved in the process and provide a meaningful explanation to all stakeholders in the process. Although the effort in the field of explainable artificial intelligence(XAI) (Gunning 2017) turned attention toward these issues, more need to be done in the process of evaluating any human facing AI systems. It has been shown that humans tend to develop deeper trust and understanding of the AI system is provided with the explanations within the human accepted conceptual realm (Weitz et al. 2019). The designers and evaluators of such AI systems and need to take into account the need for conceptual explanation provided for reasoning and behind each action, in order for human stakeholders to develop a sense of when to trust the system and when human needs to take over. Thus overcome the gap between expectation and performance of the systems.

AI explainability involves a wide range of spectrum of research. There are several stages in the AI pipeline that may provide an explanation of why a model has a certain prediction. Pre-model explainability involves a wide range of methodologies to understand the dataset involved in developing the model. Early data exploration and insight involve exploring the dataset and identifying key insight underlying concepts that may guide the AI pipeline toward better explainability in later stages. The next phase involves developing models with explainability in mind. Traditionally, this meant restricting model designer to using machine learning techniques that provide an explanation or easy to decompose into simpler problems. Due to the complex nature of the modern deep learning techniques and the sheer volume of the data, it may not be a feasible choice. The last phase involves interpreting results from already developed models in an ad-hoc fashion. Because the majority of models are developed with improving incremental improvements in pop-

ular benchmarks like Imagenet, they don't have an inherent explainability goal, a lot of effort is spent to explain results from these methods. Taking into account these phases in the AI pipeline, it is critical to incorporate procedures in the AI pipeline to provide conceptual explanations for each step in the process.

In the example of the AI tutor system (Paudyal et al. 2019) which is designed as a modular combination of various subsystems, that can be used to provide a conceptual human-centered explanation. The explanation can be generated by using the weights on the linear combination. In another approach, the subsystems can give independent explanations for the components. There is no real evaluation metric that can be used to compare these types of explanations to each other. One such metric that could be determined might be the final performance of students towards learning outcomes when using each of the systems. If a large enough sample set of students can be established, these system-wide evaluations can be a proxy for evaluating subjective artifacts such as explanations. Thus lead to a system that is human-centric and provides feedback that end-user can relate to.

Although designing an AI pipeline with explainability in mind might affect some performance metrics(Gunning 2017), a wide adaptation of such systems in human-interfaces applications requires us to reconsider our best practices in AI. It needs to include the evaluation techniques favoring evaluation in terms of explainable concepts that take a human-centered approach in all phases of the AI pipeline.

BP5: Iterative Evaluation on Adaptive Human AI Interaction

A significant aspect of Human AI interaction is that human behavior changes with continued usage of an AI system. Humans are adaptive in nature and they tend to naturally modify their behavior to either assist the AI in performing collaborative tasks or to hinder to it to test the capacities. For instance, while using personal assistants, people will modify their natural patterns of conversation to facilitate correct transcription. On the other hand sometimes people will intentionally try to make the speech recognition and language understanding task more difficult to test the limits. The evaluation conditions during training do not take into account for this 'real-world' usage. In addition, the behavior modifications shown by humans will become even more pronounced with time and more interaction. In the case of the AI Sign Language Tutor system (Paudyal et al. 2019), humans might adapt to the specific ways in which the AI does the comparisons to boost their ability to 'pass' the practice in the first time. This might lead to suboptimal learning outcomes. Conversely, humans may also begin to behave in a way that minimizes errors in the AI due to extraneous movement and other unrelated factors. These circumstances cannot be captured by using training datasets that do not take into account actual human interactions. The final evaluation of the AI sign tutor was done in a cross-sectional study with 26 participants. The evaluation also included some recall and execution tests after a learning session. While this study captures an important metric which is 'progress toward the learning outcome', the long term consequences of how humans adapt to using the AI is not captured.

In another example of the AP system, the issue of phantom carbs can be attributed to the adaptive nature of human interaction. According to one of the reviews of the Medtronic 670G system, through continued usage the user realized that the AP device was conservative and more insulin should be infused to manage postprandial hyperglycemia. The AP device includes a check where the auto mode cannot be used to infuse large amounts of insulin. Hence, the user enters phantom carbs to get more insulin in a way tricking the controller.

These effects occur in runtime and are nearly impossible to evaluate during the design and testing time. However, a good evaluation mechanism should adapt to new test cases and perform an iterative requirements verification especially for those related to safety of the human user.

A potential approach can involve novel test case prediction from post market evaluation studies. The aim is to use input output traces obtained from post market evaluations to mine test cases that were not observed in the pre-market evaluations. Then predict novel test cases that may occur during future deployments through combinatorial analysis of design variables. This will help to not only improve coverage but also failures in these test cases can be better explained if not prevented. Techniques such as metamorphic relation based test case prediction can be explored. A metamorphic relation (Segura et al. 2018; Chen et al. 2016; Zhou et al. 2018), is a property satisfied by the input output variables of the AI system as a result of the intended performance. These relations are application specific and should be derived through expert guidance. They are often a result of underlying physical or control processes. The relations will then be used to cluster the test cases into equivalence classes. A representative test case can be incorporated in the iterative evaluation experiments.

Conclusions

In this paper, we discussed the necessity for the incorporation of the effects of human interaction with AI systems in its evaluation. We hypothesize a more human-centric evaluation of AI systems can potentially reduce the gap between expected and practically observed performances of AI systems. This includes transparency in evaluation methodology, explicitly addressing participant bias, using robust metrics, and providing explanations for decisions taken by the AI system. Evaluating AI systems simplistically based only on performance-centric evaluation metric might lead to lofty expectations of performance or hype. However, if a thorough consideration is done in selecting the evaluation criteria and metric, a more robust system can be engineered. In other words, by being explicitly aware of the limitations imposed by the various assumptions in evaluation more realistic expectations of AI systems can be established and the shortcomings can be mitigated. We also advocate an iterative evaluation methodology, where post-market evaluations and user experiences can be incorporated in the evaluation of the AI system.

References

Brown, S.; Raghinaru, D.; Emory, E.; and Kovatchev, B. 2018. First look at control-iq: a new-generation automated insulin delivery system. *Diabetes care* 41(12):2634–2636.

Chen, T. Y.; Kuo, F.; Ma, W.; Susilo, W.; Towey, D.; Voas, J.; and Zhou, Z. Q. 2016. Metamorphic testing for cybersecurity. *Computer* 49(6):48–55.

FDA. 2016. Pre market approval for minimed 670g system. https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/Recently-ApprovedDevices/ucm600603.htm.

Gee, A. H.; Barbieri, R.; Paydarfar, D.; and Indic, P. 2016. Predicting bradycardia in preterm infants using point process analysis of heart rate. *IEEE Transactions on Biomedical Engineering* 64(9):2300–2308.

Gerich, J. 2013. Pathogenesis and management of postprandial hyperglycemia: role of incretin-based therapies. *International journal of general medicine* 6:877.

Gunning, D. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2.

Gupta, S. K.; Banerjee, A.; and Lamrani, I. 2019. Systems and methods for hybrid automata mining from input-output traces of cyber-physical systems. US Patent App. 16/413,018.

Lamrani, I.; Banerjee, A.; and Gupta, S. K. 2018. Hymn: Mining linear hybrid automata from input output traces of cyber-physical systems. In *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*, 264–269. IEEE.

Landgraf, R. 2004. The relationship of postprandial glucose to hba1c. *diabetes/metabolism research and reviews* 20(S2):S9–S12.

Leelarathna, L., and Thabit, H. 2018. The minimed^{TM} 670g hybrid automated insulin delivery system: setting the right expectations. *Endocrine Practice* 24(7):698–700.

2018. Medtronic minimed 670g auto mode review. https://www.below-seven.com/2018/01/31/medtronic-minimed-670g-auto-mode-review.

Messer, L. H.; Forlenza, G. P.; Sherr, J. L.; Wadwa, R. P.; Buckingham, B. A.; Weinzimer, S. A.; Maahs, D. M.; and Slover, R. H. 2018. Optimizing hybrid closed-loop therapy in adolescents and emerging adults using the minimed 670g system. *Diabetes Care* 41(4):789–796.

Nabar, S.; Banerjee, A.; Gupta, S. K.; and Poovendran, R. 2011. Gem-rem: Generative model-driven resource efficient ecg monitoring in body sensor networks. In *2011 International Conference on Body Sensor Networks*, 1–6. IEEE.

Paudyal, P.; Lee, J.; Kamzin, A.; Soudki, M.; Banerjee, A.; and Gupta, S. K. 2019. Learn2sign: Explainable ai for sign language learning. In *IUI Workshops*.

Samek, W. 2019. *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer Nature.

Segura, S.; Towey, D.; Zhou, Z. Q.; and Chen, T. Y. 2018. Metamorphic testing: Testing the untestable. *IEEE Software* 1–1.

Smith, J., and Noble, H. 2014. Bias in research. *Evidence-based nursing* 17(4):100–101.

2019. Who was really at fault in fatal uber crash? here's the whole story. https://www.azcentral.com/story/news/local/tempe/2019/03/17/one-year-after-self-driving-uber-rafaela-vasquez-behind-wheel-crash-death-elaine-herzberg-tempe/1296676002/.

Weitz, K.; Schiller, D.; Schlagowski, R.; Huber, T.; and André, E. 2019. Do you trust me?: Increasing user-trust by integrating virtual agents in explainable ai interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7–9. ACM.

Zhou, Z. Q.; Sun, L.; Chen, T. Y.; and Towey, D. 2018. Metamorphic relations for enhancing system understanding and use. *IEEE Transactions on Software Engineering* 1–1.

Zigel, Y.; Cohen, A.; and Katz, A. 2000. The weighted diagnostic distortion (wdd) measure for ecg signal compression. *IEEE transactions on biomedical engineering* 47(11):1422–1430.