TuringBox: An Experimental Platform for the Evaluation of AI Systems

Ziv Epstein [*1] **Blakeley H. Payne** [*1] **Judy Hanwen Shen** [1] **Casey Jisoo Hong** [1] **Bjarke Felbo** [1]
Abhimanyu Dubey [1] **Matthew Groh** [1] **Nick Obradovich** [1] **Manuel Cebrian** [1] **Iyad Rahwan** [1]

Abstract

We introduce TuringBox, a two-sided platform to study the behavior of artificial intelligence systems via empirical black-box testing. On one side of the platform, machine learning contributors upload existing and novel algorithms to be studied scientifically by others. On the other side, machine learning examiners develop and post machine intelligence tasks designed to evaluate and characterize the outputs of algorithms. We discuss the merits of black-box testing, outline the architecture of such a platform, and describe two interactive case studies of algorithmic auditing on the platform.

1. Motivation

As the proliferation of artificial intelligence continues, algorithmic bias within learning systems has become a popular topic of scientific study (O'Neil, 2017; Friedman & Nissenbaum, 1996; Sweeney, 2013; Hannak et al., 2014). Coupled closely with the notion of algorithmic bias is the notion of algorithmic accountability, or the idea that institutions should be held responsible for the decisions of the algorithms they use, especially in instances where algorithmic bias causes harm.

The importance of algorithmic accountability has been underscored by both the machine learning community as well as government agencies. As of January 2017, the Association for Computing Machinery adopted seven principles for algorithmic transparency and accountability in its code of ethics (ACM, 2017). More recently, the European Union passed legislation which gives "data subjects" the right to contest automated decisions. This process includes the right to an explanation as to why an algorithm made the decision in question (Goodman & Flaxman, 2016). This particular legislation is set to go into effect in May 2018.

Despite the approaching May deadline for this legislation, the machine learning community faces many challenges which prevent the systematic, scalable investigation of algorithmic behavior, which we present below:

1. **Reproducibility:** The first challenge researchers face is the difficulty in replicating AI systems of interest. A recent study showed only 6% of 400 authors at two top AI conferences shared their code, implying most state of the art algorithms are unable to be replicated for further examination (Hutson, 2018).

2. **Accessibility:** The second challenge researchers face is the increasing opacity of AI systems. Due to the use of proprietary training data or even the proprietary nature of many commercial algorithms, it is often difficult for computer scientists to access the underlying models of a system. These issues often render learning systems as "black-boxes."

3. **Efficiency:** The third challenge researchers face is a problem of inefficiency. Due to the difficulty to reproduce or access important AI systems, researchers who wish to audit learning systems are often relegated to studying a small number of systems (for example, one computer vision API) as opposed to a class of systems (such as all commercial computer vision APIs).

In this paper we discuss the merits of framing AI systems as black-boxes in order to understand their behavior. We then introduce TuringBox, a mechanism to face these challenges by allowing researchers to systematically and empirically evaluate the behavior of black-box algorithms at scale.

2. Towards a Black-Box Science

Due to their ubiquity and potential harm, the increasingly emergent and often unintended properties of these AI systems have garnered widespread attention in both the public and academic spheres (O'Neil, 2017; Friedman & Nissenbaum, 1996). However, practices such as training on proprietary data and using complex models often make it

[*]Equal contribution [1]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Manuel Cebrian <cebrian@mit.edu>, Iyad Rahwan <irahwan@mit.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

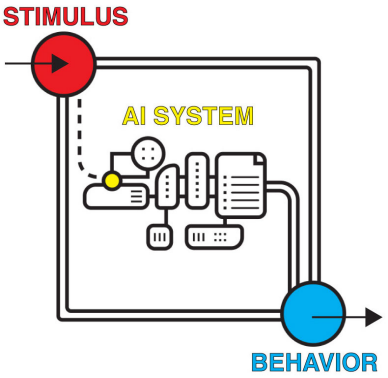

Figure 1. Conceptual illustration of applying the scientific method of experimentation and causal inference to a black box AI system (center) by providing a stimulus (left) and measuring behavioral output (right).

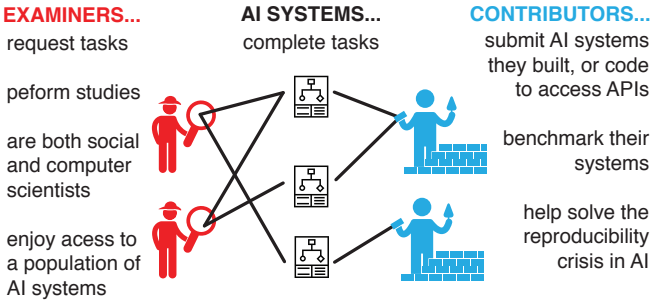

Figure 2. A schematic of the marketplace for AI.

challenging to use the underlying structure of the system to predict or study its emergent behaviors (Voosen, 2017). In order to combat this, many studies in the area of algorithmic bias have employed techniques which do not require details about the system's architecture.

In 2013, Sweeney's study on discrimination in online advertisements showed that search results were disproportionately more likely to return advertisements related to arrest when names primarily assigned to black babies were used as keywords (Sweeney, 2013). This study and others like it began a new wave of hypothesis-driven science related to AI systems.

Since then, an increasing number of studies has attempted to characterize the emergent behaviors of high-stakes algorithms. A recent study showed that darker, female faces are misgendered at higher rates than lighter, male faces by commercial facial recognition algorithms (Buolamwini & Gebru, 2018). ProPublica showed evidence of racial discrimination in new recidivism risk score algorithms as well as price discrimination based on zip code for auto insurance premiums (Larson et al., 2016; 2017). Studies have even explored algorithmic bias on deployed platforms. For example, recent studies have investigated price discrimination on e-commerce sites (Chen et al., 2016; Hannak et al., 2014).

We propose framing AI systems as black-boxes, and their output as behavior, with its own patterns and ecologies (see Figure 1). To this end, we propose using scientific techniques like experimentation and causal inference to understand these behaviors, agnostic to the underlying system architecture. We believe this framing is a first step towards understanding algorithmic behavior and increasing accountability within the machine learning community. Each of the aforementioned studies contain three core components: an AI system contained in a controlled environment, systematic stimuli, and a measure of behavior, as shown in Table 1.

3. System Overview

To assist with the scaled auditing of black-box AI systems, we introduce TuringBox. The schematic of the TuringBox framework is shown in Figure 2. On one side of the platform, *contributors* upload algorithms in various forms. The contributor community may consist of researchers, companies, or machine learning hobbyists. First, contributors will upload implementations of existing AI systems. Second, they will upload their own novel AI systems. Third, contributors will upload scripts they developed which access APIs. To encourage uploads, TuringBox eases the benchmarking process of conducting AI research by automatically comparing a contributor's uploaded algorithm to other algorithms on the platform with respect to accuracy, fairness, or other metrics determined by the examiners. For each upload, contributors will gain reputation points on the platform as a function of the algorithm's performance in these categories.

On the other side of the platform, *examiners* investigate the output of algorithms. As with previous investigations, these algorithms are studied agnostic to underlying system architecture via empirical black-box testing, and are thus represented only by their inputs and outputs (Larson et al., 2016; 2017; Buolamwini & Gebru, 2018).

These investigations may take one of two forms. First, an examiner can browse the platform for existing algorithms to be studied. An examiner will provide input data to the selected algorithms, which are specified by structured input and output, and then analyze the output of the algorithms. Second, an examiner can post a *machine intelligence task*, which calls for the creation of new algorithms by the contributor side of the platform. Examiners also receive reputation points as a function of the quality of their studies, as determined by their peers as well as other contributors.

Table 1. Sample of existing behavioral studies of AI in terms of their stimulus, AI system, and measured behavior. We describe the treatment types for each stimuli and the level at which the study occurs (either locally, via an API, or "in the field").

Study	Stimulus	Treatment Groups	AI System	Scope	Metric
Sweeney et al., 2013	Names as Search Terms	Racial Association	Google Search Engine	Via API	Disparate Treatment
Buolamwini et al., 2018	Parliamentarian Headshots	Gender, Fitzpatrick Skin Type Class	Facial Recognition Algorithms	Via API	Disparate Mistreatment
Hannak et al., 2014	Consumer Profiles	Web Broswer, Operating System, User History	Online Pricing Algorithms	Field	Disparate Treatment
Kay et al., 2015	Topics as Search Terms	N/A	Google Image Search	Field	Representat-iveness

3.1. Technical Architecture

Contributors upload Python files to the platform, which integrates them into the codebase after verifying the script matches the criterion of a well-structured algorithm.

Examiners interact with the platform via a GUI. Using the interface, they can specify the algorithms they want to access, and the dataset they wish to input. The request is then sent to the server, which returns results after server-side computation has terminated.

The system uses a virtualization technology to enforce a fine-grained security policy during computation. This ensures that the uploaded algorithms are performing the advertised computation and are not abusing the cloud computation environment.

4. Workshop

In this workshop, users will have the opportunity to perform two experiments on the TuringBox platform from both the contributor and examiner perspectives. These case studies have been carefully selected to represent two important domains in machine learning: computer vision and natural language processing.

For each case study, we will provide the user with learning systems and datasets to complete the experiment. As a contributor, the user will be able to upload their own system and use TuringBox's automatic benchmarking tool. As an examiner, users will be able to upload a customized dataset to the platform and select which algorithms to test. Users will then be able to see the output for each algorithm selected for the specified input data and have an opportunity to examine the results. We describe each case study below.

4.1. Case Study 1: Disparate Treatment by Body Type in Commercial Computer Vision APIs

Recently, computer vision systems have been shown to exhibit racial and gender biases (Buolamwini & Gebru, 2018). In this demonstration, users will examine and quantify bias in commercial computer vision systems with respect to a previously unexamined source of bias: body type.

In order to quantify bias, users will be able to select which commercial computer vision algorithms they would like to test and what outputs (for example: text labels, raciness scores, demographic information, etc.) each algorithm should return per image. The user will also specify which data set they would like to input to their selected algorithms. We will provide a dataset of images labeled for body type as a default input data set, but also provide functionality for the user to upload their own, or use the webcam in real time.

The user can then analyze each algorithm as well as the entire class of commercial computer vision algorithms for evidence of disparate treatment between body types.

4.2. Case Study 2: Sentiment Classification Bias by Gender, Ethnicity, and Age in NLP systems

Recent research has highlighted corpora and dataset biases learned by NLP models and the troubling potential impact of biased NLP systems (Bolukbasi et al., 2016; Zhao et al., 2017; Blodgett et al., 2016). While most existing studies focus on characterizing and removing bias from subcomponents of NLP models, there has only been limited research of the degree to which end-to-end real-world systems exhibit problematic biases.

We provide a demonstration that allows users to easily investigate the biases of real-world systems for sentiment classification. Users enter keywords related to demographic

attributes (e.g. gender, ethnicity, age) and examine the biases that arise across different APIs. Our demonstration quantifies the amount of bias related to the keywords by measuring differences in sentiment when doing keyword replacement (e.g. replacing 'white' with 'black'). The keyword replacements are done across a large text corpus to obtain robust measures of bias. Through this demonstration, users can easily test and characterize the output of black-box NLP algorithms.

4.3. Demonstration of Algorithm Sandbox for Contributors

For both of the two case studies above, we will also demonstrate our system's ability to integrate new algorithms into the underlying codebase. This integration protocol will enforce the following constraints: first, that uploaded algorithms are well-structured for the tasks they claim to complete, and second, that the algorithms compute under the enforcement of a fine-grained security policy. For this demonstration, participants will upload local files corresponding to pre-trained neural networks. In real time, they will be able to visualize the system verifying and integrating the algorithm, as well as the system benchmarking it against other algorithms on the platform after it has been integrated.

5. Discussion

Because of AI systems' emergent complexity, their ubiquity in society, and their inherent opacity, there is a need to build tools that increase our understanding about how these systems work in the wild and support accountable development. Our examination suggests there is untapped potential for the hypothesis driven, black-box scientific investigation of learning systems. In order to keep up with the proliferation of these systems, the machine learning community must continue to innovate new ways to evaluate and compare existing systems, in addition to building new ones.

We believe a two-sided platform is crucial to reach these goals. It offers a general yet unified framework for understanding when bias occurs in complex architectures. Many studies across computer science, behavioral economics, and legal studies have already investigated the behavior of AI systems but each uses an ad hoc approach to collecting data and measuring behavior. Our platform provides a toolkit for examining behavior across a population of AI systems, which enables a scalable and flexible alternative to costly algorithmic audits. As an increasing number of AI systems are deployed every day, and their real world stakes increase, a consistent methodology to perform these algorithmic audits at scale becomes imperative. Indeed, TuringBox or similar platforms could act as third party auditing agencies and offer certifications to companies for the ethical behavior of their AI systems.

References

Statement on algorithmic transparency and accountability, 2017. URL https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Blodgett, Su Lin, Green, Lisa, and O'Connor, Brendan. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1120.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y, Saligrama, Venkatesh, and Kalai, Adam T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Buolamwini, Joy and Gebru, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Chen, Le, Mislove, Alan, and Wilson, Christo. An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the International World Wide Web Conference (WWW'16)*, Montral, Canada, Apr 2016.

Friedman, Batya and Nissenbaum, Helen. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, July 1996. ISSN 1046-8188. doi: 10.1145/230538.230561. URL http://doi.acm.org/10.1145/230538.230561.

Goodman, Bryce and Flaxman, Seth. Eu regulations on algorithmic decision-making and a "right to explanation", 2016. URL http://arxiv.org/abs/1606.08813. cite arxiv:1606.08813Comment: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.

Hannak, Aniko, Soeller, Gary, Lazer, David, Mislove, Alan, and Wilson, Christo. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the ACM Internet Measurement Conference (IMC'14)*, Vancouver, Canada, Nov 2014.

Hutson, Matthew. Artificial intelligence faces reproducibility crisis, 2018.

Larson, J, Mattu, S, Kirchner, L, and Angwin, J. How we analyzed the compas recidivism algorithm. propublica, 2016.

Larson, J, Angwin, J, Kirchner, L, and Mattu, S. How we examined racial discrimination in auto insurance prices. propublica, 2017.

O'Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

Sweeney, Latanya. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

Voosen, Paul. The ai detectives. *Science*, 357(6346):22–27, 2017. ISSN 0036-8075. doi: 10.1126/science.357.6346.22. URL http://science.sciencemag.org/content/357/6346/22.

Zhao, Jieyu, Wang, Tianlu, Yatskar, Mark, Ordonez, Vicente, and Chang, Kai-Wei. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.