IRIE
International Review of Information Ethics Vol. 28 (06/2020)

Jinnie Shin, Okan Bulut, Mark J. Gierl

Development Practices of Trusted AI Systems among Canadian Data Scientists

Abstract:

The introduction of Artificial Intelligence (AI) systems has demonstrated impeccable potential and benefits to enhance the decision-making processes in our society. However, despite the successful performance of AI systems to date, skepticism and concern remain regarding whether AI systems could form a trusting relationship with human users. Developing trusted AI systems requires careful consideration and evaluation of its reproducibility, interpretability, and fairness, which in in turn, poses increased expectations and responsibilities for data scientists. Therefore, the current study focused on understanding Canadian data scientists' self-confidence in creating trusted AI systems, while relying on their current AI system development practices.

Keywords:

Artificial Intelligence, Data Science, Explainability, Fairness, Machine Learning, Trust

Outline:

© by IRIE – all rights reserved

ISSN 1614-1687

Author(s):

Jinnie Shin, Dr. Okan Bulut, & Dr Mark J. Gierl:

- Centre for Research in Applied Measurement and Evaluation, University of Alberta
6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5 Canada
- Email: jinnie.shin@ualberta.ca; bulut@ualberta.c; mgierl@ualberta.ca

1. Introduction

Artificial Intelligence (AI) systems have become an increasingly key element of decision-making systems in our society. From a relatively simple spam detection system to a highly sophisticated self-driving car, AI has brought many surprising innovations to our lives. For example, in education, recent introduction of AI-powered systems, such as intelligent tutoring systems and automated essay graders, have drastically changed the traditional practices in classroom instruction and evaluation. According to the case studies presented by UNESCO[1], AI-powered systems in education have greatly helped to improve the use of education data to enhance the education equity and quality around the globe (Pedro et al., 2019).

While such innovations have demonstrated capacities of current AI systems with a highly successful performance with profound impact in various areas, many concerns still remain regarding the vulnerability that AI systems have exhibited. Rossi (2019) pointed out that potential exposure to bias, lack of explainability, and susceptibility to adversarial attack as key elements, which should be addressed in order for current AI system to build a trusting relationship and to be accepted by wider audiences. Similarly, UNESCO have addressed the challenges should be expected to address issues regarding inclusion and equity of AI, ethics and transparency in data collection to prepare for an AI-powered future in education (Pedro et al., 2019).

To address such concerns and broaden the applicability of AI systems, thorough and multi-faceted evaluation should be conducted on current AI systems that are more than simply evaluating the performance accuracy. Hind et al. (2018) proposed that AI systems should be able to demonstrate an increasing fairness, robustness, interpretability, and reproducibility to facilitate more trusting relationship between an AI system and human users. They also emphasized the increasing roles and responsibilities of data scientists in developing and evaluating the trusted AI systems that satisfy the common vulnerabilities and concerns around the systems.

However, regardless of the active research and the increased expectations on data scientists, Canadians report relatively low confidence in a recently conducted, large-scale survey among data scientists and machine learning practitioners. Therefore, the current study focused on understanding why Canadian data scientists report low self-confidence regarding their system development practices. We focused on understanding the development practices in terms of model fairness, robustness, interpretability, and reproducibility.

2. Literature Review

2.1. Trusted AI systems

Over the past few years, increasing numbers of leading companies have started to propose high-level principles for AI systems in order to address concerns about the current systems and to develop trusted AI systems (Rossi, 2019). For example, in 2018, Google announced seven key objectives and principles to guide and assess their AI applications. The principles highlighted the ethical beliefs of making systems that are socially beneficial, safe, and unbiased with increased accountability. Similarly, the World Economic Forum and IBM have announced serval core principles that provides comprehensive guidelines to evaluate the entire life-cycle of AI systems. More specifically, they have addressed issues regarding model reproducibility or lineage, explainability of the behaviours and decisions, and the fairness of the model as major pillars to construct trusting relationships between human users and AI systems (Hind et al., 2018).

[1] The United Nations Educational, Scientific, and Cultural Organization

2.1.1. Reproducibility or Lineage

Reproducing research findings has long been considered one of the prominent forms of validity evidence in research (Stark, 2018). If an outcome cannot be faithfully reproduced, then it is difficult to evaluate the reliability and the generalizability of the findings. In computational science, reproducibility often refers to providing enough information to make the findings replicable by readers (Stark, 2018). For example, sharing source codes have long been considered an established practice among researchers. Despite the importance of these traditions, a crisis in reproducibility has started to draw increasing attentions among AI researchers (Hutson, 2018). Talagala (2019) explained that complex model development settings, such as the large number of artifacts, algorithm settings, code versions, system parameters, and dataset, contribute to make reproducing systems more challenging. Therefore, ensuring the reproducibility of the system requires the maintenance of a precise lineage and provenance that led to the development and usage of the system (Sridhar et al., 2018). Hind et al. (2018) suggested that data scientists should pay closer attention to make their system easy to reuse and reproduce by maintaining the algorithms and datasets for testing purposes.

2.1.2. Explainability

The majority of current machine learning and AI systems are black boxes to a certain degree, which indicates the difficulty of users to inspect and understand how and what a system did to draw its conclusions. Especially with most successful AI systems being based on complex algorithms, such as deep learning, increasing the explainability of the system is considered one of the outstanding problems that AI systems have encountered (Rossi, 2019). An explainable AI system refers to a transparent system that could be easily understood and interpreted by humans regarding how the system could arrive at a specific decision (Sample, 2017). Especially in domains like education and medicine, where the inferences are as valued as accuracies of the performance, development of self-explanatory AI system is critical to broaden its applications. If the system could have the ability to self-explain their actions and decision-making processes, then users will be able to understand the rationales for the outcomes. Siau and Wang (2018) asserted that model explainability is a critical element to create initial trust and to facilitate continuing trusting relationship between human users and AI systems.

2.1.3. Fairness

Developing a fair and unbiased AI system refers to generating systems that do not amplify or take our contextual or cultural biases (IBM, 2018). Therefore, fairness is a multi-faceted concept that could not be easily defined and is highly sensitive to the users' background (Bellamy et al., 2018). For example, in the recent conference on Fairness, Accountability, and Transparency, Narayanan (2018) introduced twenty-one mathematical and theoretical definitions of fairness. Thus, although fairness is actively research area, the challenge lies in providing clear and adaptable guidelines and metrics to best address and evaluate whether the model is fair and unbiased in various scenarios (Bellamy et al., 2018). In addition, bias can enter the system in various stages of model development and evaluation, such as through training data due to unwanted to error in labelling or from sampling procedures (Hind et al., 2018). Therefore, it is critical for data scientists to explore different bias handling strategies in different model development cycles with careful considerations of their cultural and contextual background. Hind et al. (2018) proposed several considerations regarding potential bias, ethical issues, or safety risks that data scientists should be aware of in the life-cycle of AI systems. They also emphasized the importance of exploring the strategies and remediation to avoid unfair and biased modelling.

2.2. Present Study

With increasing responsibilities and expectations that data scientists encounter to develop trusted AI systems, it is critical to understand the current state of data scientists and their system development practices. In particular, we have gleaned some issues among Canadian data scientists, as they depicted relatively low levels of self-confidence, below the world average, in a large-scale data science and machine learning survey study conducted in 2018. Therefore, the purpose of the current study was two-fold. First, we attempted to understand the factors that influence Canadian data scientist's confidence. Second, we focused on understanding their current AI system development practices to provide more trusted system.

3. Methods

3.1. Dataset

We used the survey responses gathered in the machine learning and data science survey[2] held by Kaggle in 2018. Kaggle is an online community for data scientists and machine learning practitioners to share datasets publicly and ideas regarding machine learning projects. Kaggle often hosts machine learning competitions, and the survey responses used in the study was released as part of a Kaggle competition.

Six hundred and four Canadian data scientists participated in the survey, which included 50 selected-response questions followed by thirty-six free-form responses. The questionnaires included a broad scope of questions to thoroughly understand the demographic and background information of the participants (e.g., education level), their work experience (e.g., current role as a data scientists, job titles), and their common work practices. More specifically, the last ten questions in the survey aimed to understand the current model development procedures of data scientists.

For example, questions 49 and 50 specifically focused on model reproducibility-related practices, asking "What tools and methods do you use to make your work easy to reproduce?" and "What barriers prevent you from making your work even easier to reuse and reproduce?" In terms of model interpretability question 45 and 46 stated, "In what circumstances do you explore model insights and interpret your models' predictions?" and "What methods do you prefer for explaining or interpreting decisions that are made by machine learning models? Last, in terms of fairness, question 44 asked, "What do you find most difficult about ensuring that your algorithms are fair and unbiased?" In addition, one of the interesting questions was included to evaluate the participant's confidence as a data scientist, which asked "Do you consider yourself a data scientist?"

3.2. Analysis Procedures

The analysis approaches were two-fold. We used a two-step cluster analysis to understand the factors that influence Canadian data scientists' confidence. Then, we used a Bayesian network analysis to uncover the common model development practices among Canadian data scientists to create trusted AI systems in three dimensions: reproducibility, interpretability, and fairness.

3.2.1. Two-step Cluster Analysis

Prior to the clustering analysis, we recoded the variables of interest to increase the interpretability of the analysis results. We selected six variables that are related to the participants' demographic information and

[2] The dataset is openly available at https://www.kaggle.com/kaggle/kaggle-survey-2018

their work experiences. For example, we selected participants education level (1=below high school, 2= Bachelor's or equivalent, 3= Graduate degree), their job title (1= data scientists/journalists and database engineer, 0= Others), whether their employer implements machine learning at work, and the total number of roles they have at work as a data scientist, whether they use Jupyter or IPhython to compile their work, the number of hours they spend actively coding at work (0 = 0%, 1=1% to 25%, 2=25% to 50%, 3= 50% to 74%, 4= 75% to 100%), and their confidence as a data scientist (0= Definitely not, 1= Probably not, 2= May be, 3= Probably yes, 4= Definitely yes)

Two-step cluster analysis was used to understand and compare varying responses from Canadian data scientists to uncover common profiles (or patterns) among data scientists with similar confidence levels. Two-step cluster analysis is an exploratory tool that attempts to reveal natural groupings (or clusters) within a data set (Şchiopu, 2010). Unlike traditional cluster analysis, the two-step cluster analysis could handle both categorical and continuous variables using the Euclidean distance or the likelihood distance measures and automatically determine the optimal number of clusters in a given data set. The optimal number of clusters is decided based on the ratio of distance measures, which compares the ratio of model-fit changes among adjutant cluster numbers using Schwarz's Bayesian Criterion (BIC). In our study, we selected this approach to uncover patterns and trends between the responses of survey participants and evaluate the common profiles of Canadian data scientists using their education level, job title, data scientist roles, and their work characteristics.

3.2.2. Bayesian Network Analysis

The Bayesian network analysis was used to understand the dynamics of machine learning model development practices of Canadian data scientists. Prior to the analysis, we extracted the responses where we labelled extreme behaviours in model development practices in terms of model reproducibility, interpretability, and fairness. For example, under reproducibility questions, we focused on the responses where Canadian data scientists responded "I do not make my work easy for others to reproduce", "Making my work easier to reuse and reproducible is too time-consuming", and "There is no barrier that prevents me from making my work easier to reuse and reproduce." In terms of interpretability, we focused on the responses where the participants indicated that "I do not explore and interpret model insights and predictions." and "I do not use model explanation techniques." Last, for model fairness, responses were included if they indicate, "I have never performed a task to ensure that my algorithms are fair and unbiased.

All variables represented extreme model development behaviours with binary responses, thereby, making the Bayesian network analysis quite suitable for exploring the dynamics between the variables. Bayesian network analysis attempts to model conditional dependencies among the variables using a graph modelling approach. The conditional dependencies in the network is represented using edges while nodes represent random variables. For example, in our study, each statement represented a random variable under the categories of model reproducibility, interpretability, and fairness categories. Then, the dependencies or causal relationships between the variables were represented by connecting nodes with edges in a graph.

4. Results

4.1. Findings regarding the Canadian Data scientists Confidence

We identified moderate correlations among the six selected variables with one's data scientist confidence. For example, the total number of roles they have as a data scientist at work (r=0.41, n=485, p<0.001), whether they are currently employed as data scientist, journalists, or database engineers (r=0.34, n=485, p<0.001), whether their current employer implements machine learning at work (r=0.31, n=608, p<0.001), their education level (r=0.30, n=485, p<0.001), whether they use Jupyter and IPhyton (r=0.24, n=485,

p<0.001), and the amount of time they spend actively coding at work (r=0.23, n=485, p<0.001) were significantly correlated with their data scientist' confidence.

The results from the two-step cluster analysis indicated three different groups of data scientists with different profiles, with the larges ratio of distance measure, 2.485. The cluster showed a fair quality based on the Silhouette measure, 0.40, and 18 responses were flagged as outliers and removed. Silhouette measure is a common adopted metric to evaluate the consistency within the clusters by comparing how similar within cluster objects are compared to the objects in other clusters. The first profile represented a group of Canadian data scientists who showed the highest level of data scientist confidence, followed by profile two and three. Data scientists who were identified with the first profile, showed the highest number of total roles at work as a data scientist. Also, majority of the data scientists in this group were currently hired as a data scientist, journalist, or database engineer. While there was a slight difference between their level of education and the number of hours they use actively coding at work, but the differences were not statistically significant (see Figure 1 and 2). In short, the findings indicated that the total number of roles, and job title, and whether their employer implement machine learning at work, influence the level of confidence as a data scientist the most.

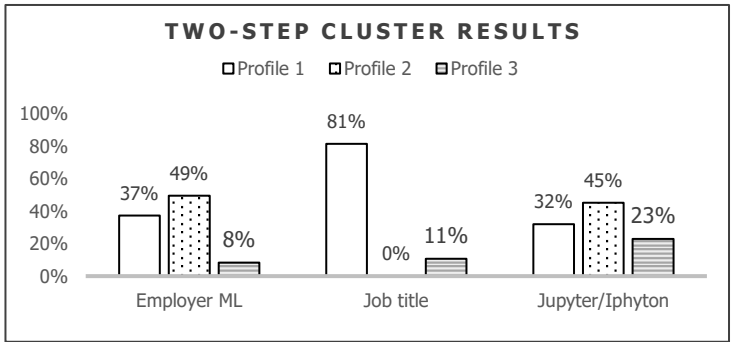

Figure 1 and 2. Two-step cluster analysis results based on three profiles.

4.2. Findings regarding the Trusted AI Systems Development Practices

About a half of Canadian data scientists indicated that making their model reproducible is too time-consuming and they have never performed a task to ensure that my algorithms are fair and unbiased, each stating "Making my work easier to reuse and reproduce is too time-consuming" and "I have never performed a task to ensure that my algorithms are fair and unbiased". Also, quite a noticeable proportion of responses indicated that they there is no barrier in reproducing their work (16%), they never explore additional strategies (12%) or attempt to make their work interpretable or explainable (9%). Each responses corresponded to the statements, "There is no barrier that prevents me from making my work

easier to reuse and reproduce", "I do not use model explanation techniques", and "I do not explore and interpret model insights and predictions."

Findings from the Bayesian network analysis could identify interesting dynamics between the variables to locate common practise regarding trusted AI system development (see Figure 3). First, Canadian data scientists who work in less time-constrained environments showed higher probability to respond that they face less or no barrier in making their model easy to reuse and reproduce. This directly led to more attempts and explorations to increase the reproducibility of their work. Positive model development practices regarding reproducibility was directly related to more explorations of skills and strategies to make their systems more interpretable and meaningful. Increased explorations of model explanation strategies were directly associated with more attempts to make the behaviour and outcomes of their system more interpretable. Last, positive practices to improve model reproducibility and interpretability led to significantly increased considerations in generation unbiased and fair model.

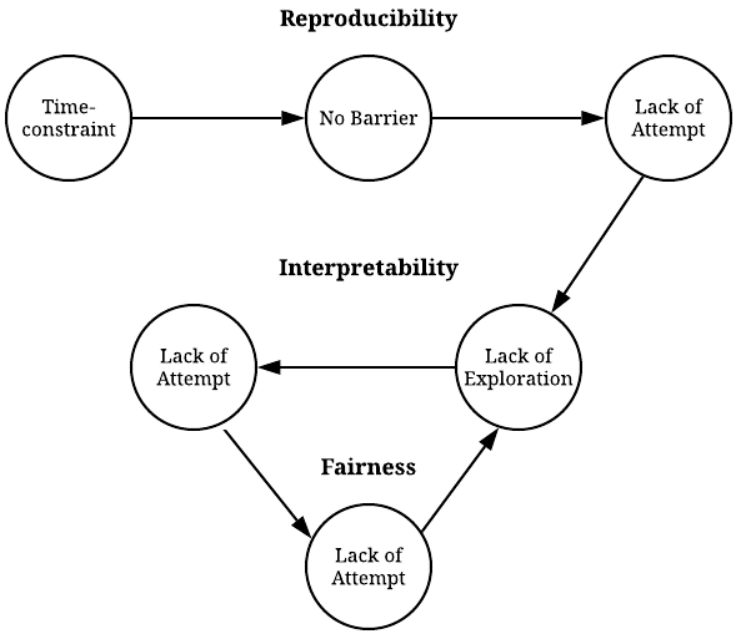

Figure 3. Conceptual representation of the Bayesian network analysis results.

5. Conclusion and Discussion

Developing an AI system that could be trusted by human users has been considered one of the key problems among AI researchers. Siau and Wang (2018) explain the increasing need in the production of higher-level AI corresponds to the movement of moving toward continuous trust development between AI systems and human users. The performance accuracies and efficiencies that early AI systems have demonstrated was a key element to facilitate the initial trust to human users. However, for the current AI systems to widen their applications to a broader audience, data scientists should provide more careful attention to develop models that are reproducible, interpretable, and unbiased (Siau & Wang, 2018; Hind et al., 2018)

Despite the increasing responsibility, a recent survey study of data scientists and machine learners have revealed that Canadian data scientists presented confidence levels that are below world average. Around one-third of the Canadian participants responded negatively when asked whether they considered themselves to be a data scientist. Therefore, the current study focused on understand the potential factors

that could impact data scientist's confidence in Canada. In addition, we attempted to understand the common AI system development practices among Canadian data scientists.

The results yielded several important implications and we could identify several distinctive model development dynamics among Canadian data scientists. First, data scientists who engaged in various number of roles at work where their employer implemented machine learning tended to show the highest level of confidence. Also, the majority of them were currently with certain job titles, such as data scientists, journalists, or a database engineer. Second, Canadian data scientists who work in less time-constrained environments showed higher probability to engage in healthier model development practices that could led to trusted-AI system development. Third, exploring model reproducibility was commonly identified as a starting point to practice more transparent and rigorous model development, which led to higher interpretability and fairness.

6. References

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., & Nagar, S. (2018). "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias". arXiv preprint arXiv:1810.01943, 2018.

Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., & Varshney, K. R. "Increasing Trust in AI Services through Supplier's Declarations of Conformity". arXiv preprint arXiv:1808.07261. 2018

Hutson, M. "Artificial intelligence faces reproducibility crisis". 2018.

IBM. "Trusted AI". Retrieved May 21, 2019, from https://www.research.ibm.com/artificial-intelligence/trusted-ai/Narayanan, Arvind. "Translation tutorial: 21 fairness definitions and their politics." In Proc. Conf. Fairness Accountability Transp., New York, USA. 2018.

Pedro, F., Subosa, M., Rivas, A., & Valverde, P. "Artificial intelligence in education: challenges and opportunities for sustainable development". Rossi, Francesca. "Building Trust in Artificial Intelligence." Journal of International Affairs 72.1 (2019): 127-134.

Sample, I. "Computer says no: why making ais fair, accountable and transparent is crucial". *The Guardian*. 2017.

Şchiopu, D. "Applying TwoStep cluster analysis for identifying bank customers' profile". *Buletinul*, 62, 2010: 66-75.

Siau, K., & Wang, W. "Building trust in artificial intelligence, machine learning, and robotics". *Cutter Business Technology Journal*, 31(2), 2018: 47-53.

Sridhar, V., Subramanian, S., Arteaga, D., Sundararaman, S., Roselli, D., & Talagala, N. "Model governance: Reducing the anarchy of production" {ML}. In 2018 {USENIX} Annual Technical Conference ({USENIX}{ATC} 2018: pp. 351-358).

Stark, P. B. "Before reproducibility must come preproducibility". Nature, 557 (7707), 2018: 613.

Talagala, N. "ML Integrity: Four Production Pillars for Trustworthy AI", 2019. Retrieved from https://www.forbes.com/sites/cognitiveworld/2019/01/29/ml-integrity-four-production-pillars-for-trustworthy-ai/#2c2d31005e6f

© by IRIE – all rights reserved

ISSN 1614-1687