The Definition of Machine Learning Interpretability and Its Impact on Smart Campus Projects.

By Ms Raghad Zenki, PhD candidate at the University of Northampton.
Dr Mu Mu, Senior Lecturer at the University of Northampton. on 24 May 2019
{raghad.zenki,mu.mu}@northampton.ac.uk

Research Motivation

Machine learning (ML) has shown increasing abilities for predictive analytics over the last decades. It is becoming ubiquitous in different fields, such as healthcare, criminal justice, finance and smart city. For instance, the University of Northampton is building a smart system with multiple layers of IoT and software-defined networks (SDN) on its new Waterside Campus. The system can be used to optimize smart buildings energy efficiency, improve the health and safety of its tenants and visitors, assist crowd management and way-finding, and improve the Internet connectivity.

With ML systems growing popular, many questions start to emerge: How much should human trust ML models? Will ML work on new "unknown" data? What can it tell about the world? Can human depend on AI partners? It is believed that the remedy of the concerns above hinges on the ability of ML/AI models to reason and interact with humans. Thus there is a need to develop models that can deliver good (performance) and explainable (interpretability) outcomes.

The Definition and Impact of ML Interpretability

Defining interpretability is a challenging task. Many researchers produce diverse motivations over the notion of interpretability and offer a myriad of features to classify a model as interpretable. However few have clearly and distinctly defined interpretability, especially its beneficiaries, shareholders, and overall significance to businesses and communities. Consequently, we can conclude that either: a) the interpretability is something globally agreed upon without being explicitly framed in words, or b) it is an inconspicuous term, and therefore, some explanations regarding the interpretability of ML models may lead to quasi-scientific claims. The author's observations of the literature imply that the latter seems to be the case. The variance and the occasional conflict, between the purpose of the interpretability and the explanation of many interpretable models, indicate that interpretability is not a monolithic concept, and it reflects a different number of diverse notions.

Linguistically, Interpret means *to explain a meaning or to present in understandable terms* [(1)]. According to ML systems, interpretability could be defined as the capability to elaborate and reason the models' behavior in a way that humans can comprehend. Nevertheless, the precise outlines of interpretation remain elusive; [1] described the explanation, in the context of psychology, as "*the currency in which human exchanged beliefs*". The questions, then, arise:

- What shapes an explanation?
- What makes an explanation understandable?
- When/how it has to be generated?

---

[1] Merriam-Webster dictionary, accessed 2019-05-20

- How ML applications interact with humans for informed decision marking?

Some research has characterised the interpretability as a means to give some sense of mechanism [2][3].

The potential influence of interpretability is multi-faceted. Interpretability is employed as a means to assert the raison d'etre of ML systems. Many requirements should be optimized in any ML systems, concepts of *Fairness* and *Quality*, reflect that the system is not discriminate against certain groups. *Privacy* would prove that confidential information in the data set is adequately protected. Other features such as *Rigor* and *Robustness* refer to the algorithm performance efficiency regarding the divergence of the input parameters. *Causality* is the ability to predict that the output will change due to some disturbance in the real system. *Usability* is about providing information that may help the users to complete a specific task. Lastly, *Trust* implies that a particular system has gained human confidence, such as aircraft collision avoidance systems. Some would argue that the research communities have classified some features such as privacy [4][5] and Reliability [6], and these formalizations followed by many rigorous studies in these areas with absolutely no need for further interpretations. Nevertheless, according to [7] "*Explanations may highlight the incompleteness*", and the claim here is that interpretability assure that the ML desiderata, as mentioned above, are equally satisfied [8].

Smart Campus Survey

Taking the notion of "Interpretability can instigate the users' trust" as a starting point, the author had conducted a survey (see Appendix) at the University of Northampton Waterside Campus, where many staff and students participated in it (Fig. 1). The purpose of the survey is to help us understand the public perception of data collecting and AI on a university campus in comparison with social media. Furthermore, we investigate how ML interpretability could reshape the general opinion about data gathering and analysis.

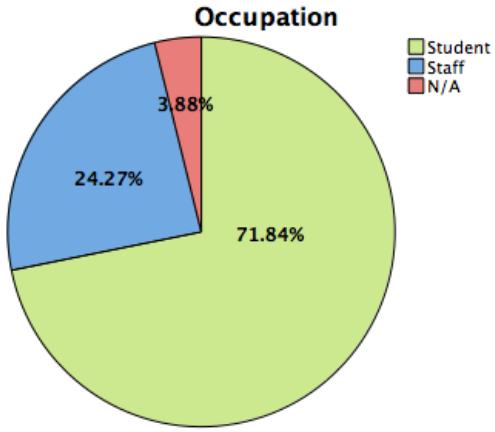

Figure 1 : participants' occupations.

We collected the users' response to the fact that their data is being collected and analyzed, on a daily basis, by companies such as Google, Facebook, and Amazon (Fig. 2).

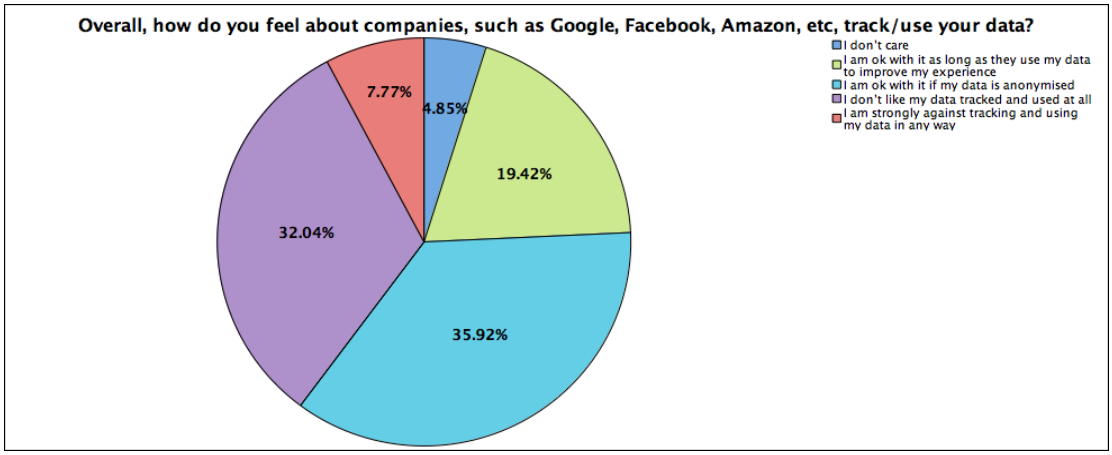

Figure 2: Shows the participants' response to their data being collected on Social Media.

Next, we asked how they would feel about their data being captured and used by the university to improve their own experience on campus (Fig. 3).

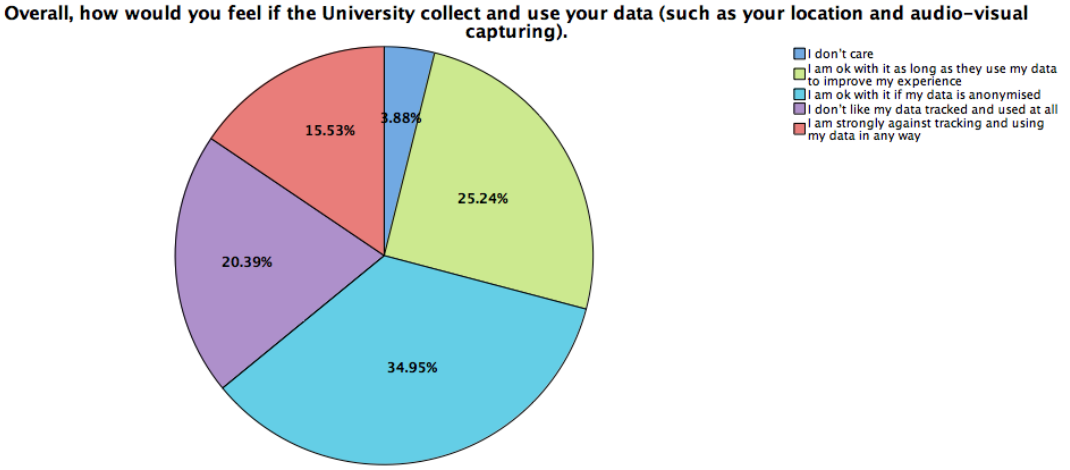

Figure 3: Shows the participants' response to their data being collected at the University of Northampton.

Then we compared the results to verify our hypothesis, in which we assumed that the staff and students would trust the university more than social media. Our data shows that 64% of the participants are okay with their data being used by the university, which is slightly higher than their response to social media (60%). To our surprise, the results were close which means that students and staff would trust the university and social media with their data almost equally (fig.4).

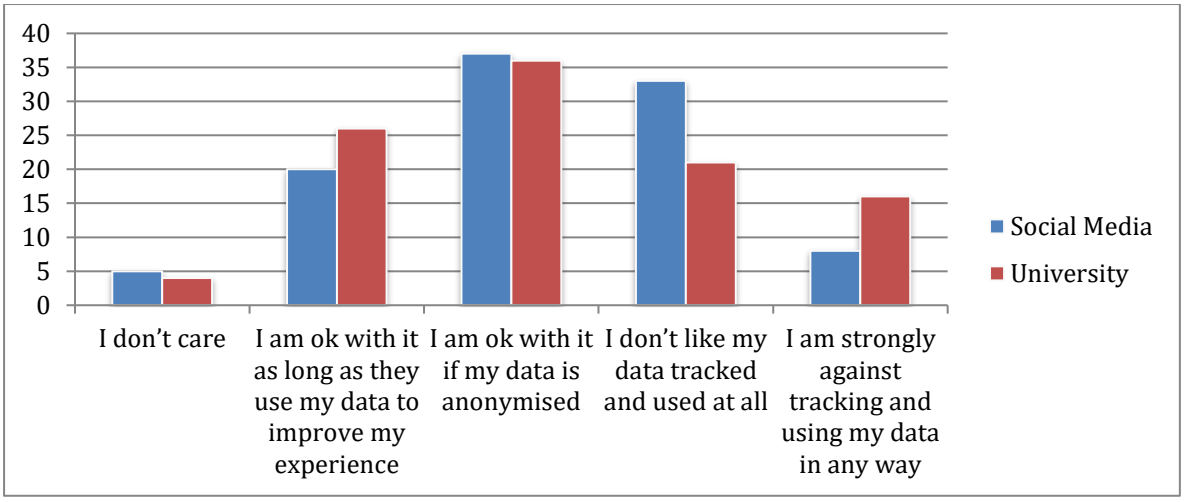

Figure 4: Shows the participants' response to their data being collected by the University of Northampton vs. Social Media.

Then we focused on the people who are on the other side of the spectrum where there are 36% of the participants against or strongly against their data being used in any format for any reason. When asked "if the University's Artificial Intelligence can explain to you how your data is used, how the decisions are made and let you make changes, would it change your opinion?", over 45% of this group of participants feel positive or very positive of the interpretable AI. (Fig.5). Looking at the opinions from all participants, around 55% of the participants responded positive or very positive to ML interpretability.

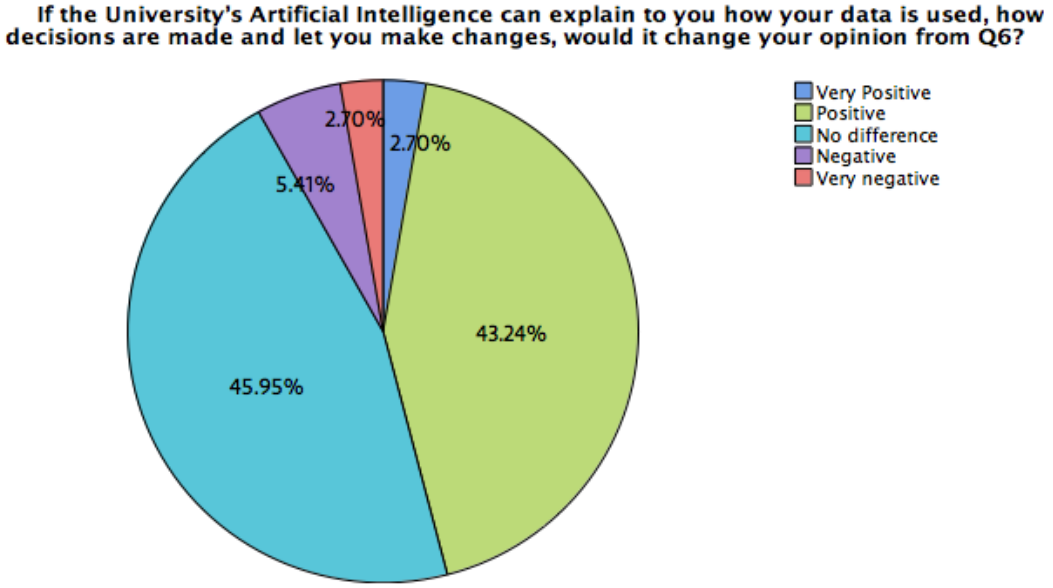

Figure 5: Shows the 36% of the participants who are against or strongly against their data being used response to using Interpretable ML model on campus.

Conclusions and Future Work

In conclusion, outlining the definition of the interpretability will lead to a better understanding of the fundamental purposes behind interpretable models. Consequentially, leading us to believe that interpretability in ML is the optimum approach to investigate, if we are to build an IoT-based model that can help the university understand how people use campus facilities, which leads to develop the university services, and enhance the students/staff experience on

campus. Our initial data show that young university students find data gathering and AI-assisted smart campus generally acceptable while over half of the survey participants have positive or very positive opinions towards ML interpretability. Moreover, in future work, we are planning to build an interpretable smart system that reasons its decisions using information visualization. Students will have the chance to provide feedback to the research team. In order to study how interpretability can bring positive changes to the perception of data-driven smart campus, we plan to conduct the survey again after deploying the system on campus.

Acknowledgement

This work is supported by UK Research and Innovation (UKRI) under EPSRC Grant EP/P033202/1 (SDCN).

References

1. Tania Lombrozo. The structure and function of explanations. Trends in cognitive sciences, 10(10): 464–470, 2006.
2. William Bechtel and Adele Abrahamsen. Explanation: A mechanist alternative. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 2005
3. Nick Chater and Mike Oaksford. Speculations on human causal learning and reasoning. Information sampling and adaptive cognition, 2006.
4. Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. Adnostic: Privacy preserving targeted advertising. 2010.
5. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science Conference. ACM, 2012.
6. Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, 2016.
7. Frank Keil, Leonid Rozenblit, and Candice Mills. What lies beneath? understanding the limits of understanding. Thinking and seeing: Visual metacognition in adults and children, 2004.
8. Doshi-Velez, F. and Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608v2, 2017.

Appendix:

Public Opinion Survey

I am part of a University research team doing research on smart campus and artificial intelligence. Please help our research by answering the following questions. Thank you!

1. Please select your gender.
   - O Male
   - O Female
   - O Other
   - O Prefer not to say

2. Age group (please circle)
   <18 18-24 25-34 35-44 45-54 >54

3. What of the following best describe you?
   - O Student
   - O Staff

4. Please describe your level of knowledge to the following on a scale.

#Option	Never heard of it	Have little idea	Have a clear idea	Can explain how it works	Expert
Artificial Intelligence	O	O	O	O	O
Smart Campus	O	O	O	O	O

5. Overall, how do you feel about companies, such as Google, Facebook, Amazon, etc, track/use your data?
   a) I don't care.
   b) I am ok with it as long as they use my data to improve my experience.
   c) I am ok with it if my data is anonymised.
   d) I don't like my data tracked and used at all.
   e) I am strongly against tracking and using my data in any way.

6. Overall, how would you feel if the University collect and use your data (such as your location and audio-visual capturing).
   a) I don't care.
   b) I am ok with it as long as they use my data to improve my experience.
   c) I am ok with it if my data is anonymised.
   d) I don't like my data tracked and used at all.
   e) I am strongly against tracking and using my data in any way.

7. In your opinion what of the following Smart Campus / Artificial Intelligence use-cases worth to be investigated?
   a) Smart building and energy efficiency (e.g., turn lights off when no one is around)
   b) Campus health and safety (e.g., detect violence and danger)
   c) Crowd management and way-finding (e.g., find quiet space or best bus to take)
   d) Improving the internet connectivity
   e) Other Please specify: _____

8. If the University's Artificial Intelligence can explain to you how your data is used, how the decisions are made and let you make changes, would it change your opinion from Q6?
   a) Very positive (I am more willing to allow the use of my data)
   b) Positive
   c) No difference
   d) Negative
   e) Very negative (I am less willing to allow the use of my data)

9. (Optional) Regarding the Smart Campus use-cases mentioned in No7 and based on your own experience in campus is there other services you would like to be included?