Explanations Considered Harmful? User Interactions with Machine Learning Systems

Simone Stumpf
Adrian Bussone
Dympna O'Sullivan
Centre for Human Computer
Interaction Design
School for Mathematics, Computer
Science and Engineering
City University London
London, UK, EC1V 0HB
Simone.Stumpf.1@city.ac.uk

Abstract

It has been suggested that the intelligibility of machine learning system behavior is an important factor in ensuring that users can identify that the system has erred, understand how the system operates and that thereby they are better able to provide appropriate feedback to the machine learning system to improve its accuracy. There has been increasing research into how to make machine learning intelligible to users without a background in AI, and it has been shown that providing explanations of a system's reasoning has many benefits. In this paper we review recent work in this area but also point to instances when explanations might have less desirable effects. Further work is warranted to understand how best to expose the reasoning of machine learning systems to improve their usability.

Author Keywords

Machine learning; explanations; reliability; intelligibility.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

Systems that rely on knowledge derived from analyzing data using machine learning algorithms, such as Clinical Decision Support Systems, or that personalize themselves to users, such as music recommender systems, are becoming widely used. While these kinds of systems are usually fairly reliable, they will make mistakes. It has been suggested that the intelligibility of a machine learning system's behavior is an important factor in ensuring that users can identify that the system has erred, understand how the system operates so that they are able to provide appropriate feedback to the machine learning system to improve its accuracy. We review some work in this area but also highlight instances when explaining the system's reasoning can be problematic.

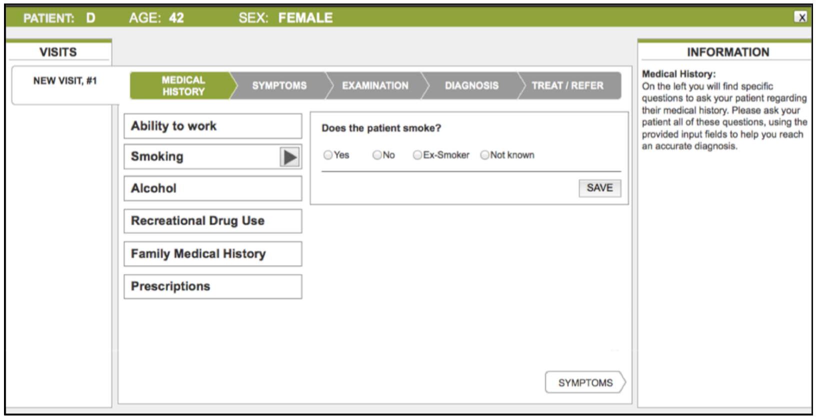
Making Machine Learning Intelligible

In the past, machine learning systems have not made their reasoning transparent to users but this is starting to change [6, 5]. Explaining the system's behavior to increase its intelligibility has been shown to have many benefits, including increased understanding of how a machine learning system operates [8, 11, 4, 7, 13], improved interactive feedback to machine learning systems [9, 16], and better user satisfaction, usability and trust in the system’s suggestions [2, 15, 3].

Previous work has suggested *what* to explain, such as inputs, outputs, and the model underlying the system's reasoning [16, 12, 11]. This can be achieved through a number of explanation types (e.g. covering questions such as What? Why? Why Not? What If? How to?) [12]. Similarly, there has been research to determine *how* to explain the system's reasoning best [9], establishing principles for "explanatory debugging" in terms of explainability (Be iterative. Be sound. Be complete. Don’t overwhelm.) and correctability (Be actionable. Be reversible. Always honor feedback. Incremental changes matter.). It has also been investigated *how much* to explain [8], showing that in some situations comprehensive explanations about inputs can be traded off against aligning soundly with the underlying machine learning model.

Explanations Considered Harmful: A Study

Previous research has shown that users do not always know how reliable an intelligent system is, and their trust might be misplaced [10]. This can result in *misuse* of the system through over-reliance on the system (i.e. the user agrees with *incorrect* system suggestions) or *disuse* (i.e. the user does not follow *correct* suggestions) [14]. Explanations can possibly compound over-reliance because they can increase user trust in the system's reliability [7], and thus cause users to trust a system when inappropriate. Our recent work [1] has started to investigate the effects of intelligent system explanations on misuse and disuse. We provide an overview of the study and results here, and discuss the implications of this work on future research directions.



PATIENT: D AGE: 42 SEX: FEMALE

Does the patient smoke?

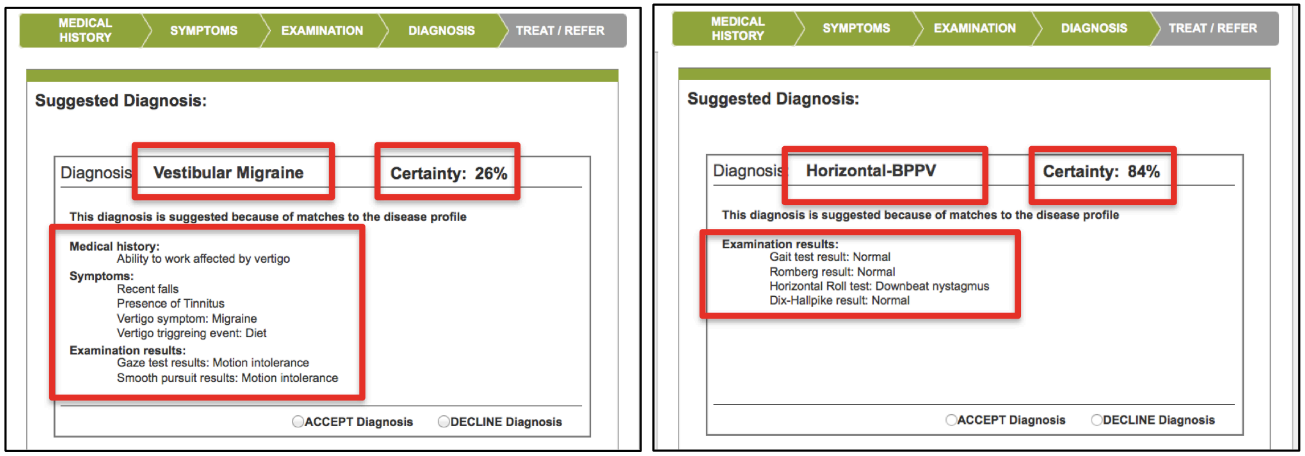
○ Yes ○ No ○ Ex-Smoker ○ Not known

Medical History:
On the left you will find specific questions to ask your patient regarding their medical history. Please ask your patient all of these questions, using the provided input fields to help you reach an accurate diagnosis.

Figure 1: The prototype. The participants entered case details into the prototype for the system to make a suggested diagnosis.

We developed a Clinical Decision Support System (CDSS) prototype within an EU-funded project (http://www.embalance.eu/) which supports primary care physicians to diagnose and treat balance disorders. In our study, we used a Wizard-of-Oz approach in which the behavior of the mocked-up prototype was controlled by the researcher, unbeknown to the participant (Figure 1). In order to simulate the experience of diagnosing patients with balance-related complaints, we created eight clinical cases. Each clinical case described a fictitious patient's age and gender, their medical history, symptoms, and the results of four clinical examinations. We adapted several aspects in the prototype to investigate explanations and their effects on reliance (Figure 2): the correctness of the diagnosis (4 correct diagnoses and 4 incorrect ones), the confidence of the system in the diagnosis shown to the user (4 high and 4 low), and the extent of the explanations given to the user (Comprehensive or Selective). We counter-balanced correctness and confidence across the 8 cases. Participants either were shown Comprehensive or Selective explanations in a between-group study design; four participants viewed Comprehensive explanations while three used Selective ones. Each participant was asked to consider all eight cases; one participant was only able to complete four. This resulted in a total of 52 cases considered altogether: 28 by the Comprehensive group and 24 by the Selective group. All participants were primary care physicians or healthcare professionals with an average of 6.5 years experience.

Our results showed that participants in the Comprehensive group agreed with more suggestions than the Selective Group, and also seemed to agree with more incorrect ones (Figure 3, red checkmarks). Thus, a larger amount of information presented in the explanation seemed to matter in *agreeing* with



MEDICAL HISTORY | SYMPTOMS | EXAMINATION | DIAGNOSIS | TREAT / REFER

Suggested Diagnosis:

Diagnosis **Vestibular Migraine** **Certainty: 26%**

This diagnosis is suggested because of matches to the disease profile

Medical history:
Ability to work affected by vertigo
Symptoms:
Recent falls
Presence of Tinnitus
Vertigo symptom: Migraine
Vertigo triggreing event: Diet
Examination results:
Gaze test results: Motion intolerance
Smooth pursuit results: Motion intolerance

○ACCEPT Diagnosis ○DECLINE Diagnosis

MEDICAL HISTORY | SYMPTOMS | EXAMINATION | DIAGNOSIS | TREAT / REFER

Suggested Diagnosis:

Diagnosis **Horizontal-BPPV** **Certainty: 84%**

This diagnosis is suggested because of matches to the disease profile

Examination results:
Gait test result: Normal
Romberg result: Normal
Horizontal Roll test: Downbeat nystagmus
Dix-Hallpike result: Normal

○ACCEPT Diagnosis ○DECLINE Diagnosis

Figure 2: The Comprehensive version (left) provided an explanation that showed all inputs associated with a diagnosis, whereas the Selective version (right) showed only inputs from examinations. We also manipulated the correctness of the diagnosis and the system certainty.

incorrect suggestions, i.e. participants misused system suggestions. A possible reason for this over-reliance was that the participants receiving Comprehensive explanations were exposed to additional justifications, persuading them to go along with the system even though they knew that the system sometimes erred. The verbalizations of participants show the persuasive nature of the Comprehensive explanation, disregarding their own diagnostic hypothesis and agreeing with an incorrect suggestion: *"I guess this thing knows more than me. The system knows more than me. I'll accept [the diagnosis]."* [C02]

On the other hand, nearly one third of the decisions made by the Selective group were disagreements, including three with correctly suggested diagnoses (Figure 3, red crosses). This suggests that showing less

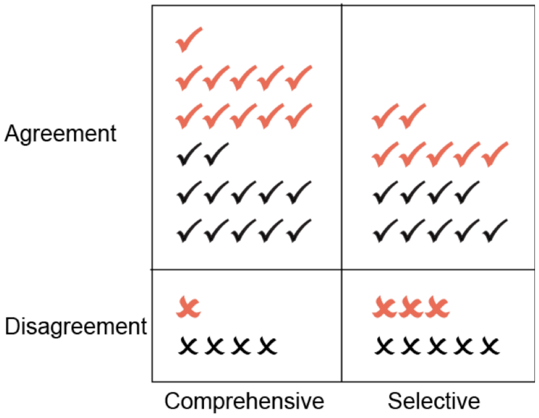


Figure 3: Number of agreements (top) and disagreements (bottom). Incorrect decisions made by participants are shown in red.

information in the explanations caused unwarranted self-reliance, that is, if not given enough information a user may choose to rely on their own limited knowledge rather than that of a CDSS.

Discussion and Conclusion

Although our study was small, our results provide preliminary results of unintentional and possibly harmful effects of explanations in machine learning systems. First, our results indicate that explanations have effects on reliance: a more detailed explanation may promote over-reliance but without providing explanations there is a danger that users will rely too much on themselves. More work is needed to establish the impact of explanations on reliance. Second, our findings also indicate explanations' complex effects on a user's trust in a CDSS. Because CDSS users who trust the system highly are also likely to over-rely on the system's suggestions, explanations need to be designed so as to carefully instill *appropriate* trust. We are interested in discussing work that can lead to intelligible and usable machine learning systems.

References

1. Adrian Bussone, Simone Stumpf, Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In Proc. International Conference on Healthcare Informatics (ICHI). IEEE, 160-169
2. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (November 2008), 455-496.
3. Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.* 58, 6 (June 2003), 697-718.
4. Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 147-156.
5. Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 497-530.
6. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (CSCW '00). ACM, New York, NY, USA, 241-250.
7. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 1-10.
8. Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing* (VLHCC '10). IEEE Computer Society, Washington, DC, USA, 41-48.
9. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the 20th International Conference on Intelligent User

Interfaces (IUI '15). ACM, New York, NY, USA, 126-137.

10. John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*, 1, 50-80.

11. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09). ACM, New York, NY, USA, 2119-2128.

12. Brian Y. Lim and Anind K. Dey. 2010. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*(UbiComp '10). ACM, New York, NY, USA, 13-22.

13. Dympna O'Sullivan, Paolo Fraccaro, Ewart Carson, and Peter Weller. 2014. Decision time for clinical decision support systems. *Clinical Medicine* 14, 4, 338-341.

14. Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2, 230-253.

15. Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (October 2012), 399-439.

16. Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *Int. J. Hum.-Comput. Stud.* 67, 8 (August 2009), 639-662.