Perceptions of Machine Learning

The genie in the bottle

Mahboobeh Harandi[1], Kevin Crowston[1], Corey Jackson[1], and Carsten Oesterlund[1]

Syracuse University, Syracuse NY 13244, USA
{mharandi,crowston,cjacks04,costerlu}@syr.edu

Abstract. We explore how people developing or using a system with a machine-learning (ML) component come to understand the capabilities and challenges of ML. We draw on the social construction of technology (SCOT) tradition to frame our analysis of interviews and discussion board posts involving designers and users of a ML-supported citizen-science crowdsourcing project named Gravity Spy. We extend SCOT by anchoring our investigations in the different uses of the technology. We find that the type of understandings achieved by groups having less interaction with the technology is shaped more by outside influences and less by the specifics of the system and its role in the project. This initial understanding of how different participants understand and engage with ML point to challenges that need to be overcome to help participants deal with the opaque position ML often hold in a work system.

Keywords: Machine learning · User perceptions · Social construction of technology

1 Introduction

Machine learning (ML) has recently increased in capability and is being more broadly applied. However, the technique has distinctive characteristics that are unlike other approaches for supporting or automating work, e.g., being trained rather than programmed, and thus dependent on the quantity and quality of data; being probabilistic rather than deterministic; and often being opaque, i.e., unable to explain decisions. These differences can cause problems for use and users. The application of an ML system is clearly an algorithmic phenomenon, but our ability to control the technology is limited: e.g., an unwanted behavior is harder to fix if it is the result of a training dataset rather than algorithm design. Given these differences, we are interested in how people, particularly non-experts, make sense of ML. The question we explore in this paper is:

> how do people developing or using an ML system realize the distinctive characteristics and limitations of the technology?

We explore this question in the context of an online citizen-science project called Gravity Spy (http://www.gravityspy.org) that incorporates ML and involves a number of different groups having varied interactions with the technology, thus providing a diversity of perspectives on ML.

2 Theory

We build our exploration on how people approach, work with and perceive ML on two basic concepts, interpretive flexibility and relevant social groups, as discussed in the Social Construction of Technology (SCOT) program [9]. Emerging out of the "Bath school" in Science and Technology Studies by Collins [5] and his students, Trevor Pinch and David Travis, this human-centered approach is concerned about the design actions taken by different groups and the meanings these impart.

First, the notion of interpretive flexibility highlights that technologies and artifacts can be associated with more than one meaning. Sufficiently underdetermined, technological artifacts allow for multiple interpretations and possible designs. Concerned with the social shaping of science and technology, Collins [5] and later Pinch and Bijker [9] suggested that technology design is an open process with different potential outcomes depending on the social circumstances of its development.

Second, the concept of relevant social group embodies people with a common interpretation, that is, all members of a certain social group that share the same set of meanings attached to a specific artifact [9, p. 414]. To determine who falls into such a group, Pinch and Bijker [9] ask a series of questions.

- First, does the artifact have any meaning to the members of the social group under investigation? Obvious groups would include users or consumers of an artifacts but one might find less obvious groups.
- Second, is a previously-defined social group homogeneous when it comes to the meanings given to an artifact or would it be helpful to break up a heterogeneous group into several sub-groups?
- Third, in defining relevant social groups, Pinch and Bijker [9] are particularly interested in the problems facing each group in relation to the artifact.
- Finally, a number of technological solutions might emerge around each problem.

By focusing on problems and solutions, Pinch and Bijker [9] do not go into details about the type of practices associated with the artifact and how groups may engage with an artifact in radically different ways, though this seems to be behind the perception of problems.

Finally, in Pinch and Bijker [9], interpretive flexibility is eventually overtaken by processes of closure and stabilization. However, as we are studying a technology as it is newly deployed, we do not expect to see this part of the process.

In summary, the SCOT approach to technology suggests identifying the relevant social groups around a technology by looking for groups with relatively homogeneous perceptions of the problems with a technology and the solutions for those problems. We extend this approach by first considering how the groups may differ in how they interact with the technology that lead to perception of problems, as well as the resources they can draw on to develop their understandings of solutions.

3 The Gravity Spy System

We examine perceptions and source of perceptions of ML technology in the context of a citizen science project called Gravity Spy. Citizen science describes an arrangement where members of the public contribute to scientific research [2]. Gravity Spy supports research in the Laser Interferometer Gravitational-Wave Observatory (LIGO) Scientific Collaboration, a consortium of researchers and institutions working to record evidence of gravitational waves [10, 1]. The LIGO uses detectors that in addition to potential gravitational waves record internal and external noise (called "glitches") generated as a result of the sensitivity that is required to recored gravitational waves. Since there are hundreds or thousands of glitches every day, scientists ask citizens to help classify glitches on Gravity Spy system. Volunteers also try to find new classes of glitches, i.e., collections of glitches with similar appearances that do not fit a known class.

3.1 Gravity spy as a hybrid human-machine system

Gravity Spy incorporates ML in three ways.

- First, a deep learning ML classifier was trained on gold data (i.e., gltches classified by experts). The ML classifies glitches added to the system into one of twenty-two known glitch classes or "none of the above". It also provides the likelihood of the glitch belong to each of the classes. The classifications are used to route gitches to volunteers, with beginners getting glitches for which the ML is more confident and more advanced users glitches with lower confidence that are presumably harder to classify.
- Second, ML is applied to support the process of finding new glitch classes. A similarity-search tools allow volunteers to search for glitches similar to a seed glitch.
- Lastly, a clustering tool identifies similar "none of the above" glitches to propose new glitch classes.

The result is a hybrid human-machine system, using ML techniques intertwined with the dataset that has been used to make a predictive model for labeling unseen data. Since the training dataset for Gravity Spy was created by the science team, their interpretations and biases affect the process of agreement and quality of the training dataset. The quality of the training data in turn affects the process of feature selection, feature extraction and the ML algorithm's prediction. Further, each group of people have a different interpretations of how the ML algorithm has classified unseen data, as there is no way to understand why it has predicted a specific result (the details are in the interaction of hidden layers in neural networks). Each group faces different problems depending on their interpretations and interactions and seeks different solutions to address the problems.

4 Research Methodology

A qualitative approach was adopted to understand how individuals approach, work with and perceive ML. Qualitative approaches have proven valuable in understanding the social and cultural significance people impart on technologies [8, 7, 6]

4.1 Data elicitation

The empirical data for our study come from two sources interviews and the Gravity Spy discussion forum posts. We conducted six **interviews**: two with volunteers who we refer to as Brandon and Katie, and four with members of the Gravity Spy science team, referred to as Peter, Casper, Marsha, and April. The selection of interviewees was based on a purposive and opportunistic sampling procedure. From the volunteer population, we chose to interview volunteers who had been a part of the project for a periods of time allowing them to have come into contact with most ML components. The goal of the interviews was to understand how an interviewees perceive ML in Gravity Spy. Using a semi-structured interview protocol we asked questions such as "Can you describe the functioning of the ML in the project and what role it plays in various stages of the work process?" Each interview was audio recorded and lasted approximately one hour and was transcribed.

As for the second source, **discussions** in Gravity Spy cover a variety of topics written by volunteers and the Gravity Spy team. We collected comments (N = 425) posted to the discussion fora pertaining to the ML functions, use, and perceptions of problems and solutions. To find relevant posts, we conducted a keyword search on the Gravity Spy homepage to search for conversation threads related to ML. We broadened our search to include related terms such as: algorithm, machine learning, pattern recognition, machine teaching, computer learning, and artificial intelligence.

4.2 Data analysis

The data were analyzed using thematic analysis [4, 3] with SCOT as a sensitizing device. We started with SCOT concepts of relevant social groups and interpretive flexibility. Once the interviews were completed, the authors read through the interview notes and transcripts identifying patterns of use, work with, and perceptions of ML. Interviewee statements were captured and organized based on similarities in how they work with the ML. We reviewed each category and developed themes around how individuals used the ML and the problems they experienced in their work. Individuals with common problems were then linked to the relevant social group. The themes that describe how they use the technology, what problems they have and how they solve problems are described in the results. We used the discussion posts to corroborate expert volunteer accounts.

5 Results

5.1 Use of the ML Classifier

As noted above, ML is currently being used in three ways in Gravity Spy. We examine one of these technologies in this paper, the ML classifier. Because of space limitations, we will not discuss the search tool and the clustering algorithm.

We identified different social groups around the ML classifier by considering first how each group use it in their work. In doing so, we followed the SCOT strategy of starting with pre-existing social groups (such as the science team) and breaking them up if they had different uses or perceived different problems or solutions.

The first group we identified are the **ML developers**, those who designed and developed the ML classifier with the aim of classifying images to known classes. They explained that they trained the ML classifier based on the gold data that was created by the LIGO scientists. Peter is identified in this group, emphasized the importance of gold dataset as:

> The basic step in all kinds of ML algorithms is having a labeled dataset. We train the ML based on the labeled data in gold dataset. ML cannot do a magic. ... The heart of ML algorithms is the gold datasets.

The developers believe that the ML classifier works well for most of the known classes. However, they need volunteers to check the results of ML classifier.

The second group are the **platform developers**, the developers who did the development work to create the Gravity Spy system. One of the important tasks they did for the project was developing the system to convert raw data about glitches provided by LIGO to images that are perceivable for volunteers and Gravity Spy team and also used by the ML classifier. Casper and Marsh are identified in this group. Casper said:

> I think the best thing we did with the machine learning from the volunteers' perspective and the LIGO perspective, is we presented the output in a really nicely digestible way, as images. People can just understand that better than the raw data. And ML also has an output in a grounded way, images.

Second, in collaboration with the LIGO scientists (see below), they classified glitches to create the initial version of the gold dataset. Third, they collaborated with the ML developers to integrate the ML classifier results into the platform. They designed the platform to assign images to different workflow based on its confidence score. However, for retiring images, they consider that the volunteers' classifications are reliable and so see a need for volunteers to check the result of the ML classifier.

The third group are the **LIGO scientists** who are the intended users of the data from Gravity Spy. Later in the project, they collaborated on fixing problems with the gold dataset. They know how improving the gold dataset and the ML classifier had a positive impact on the results of the classifications. However, they

otherwise do not know the details of the ML classifier. These scientists asked for and were given access to the results of the ML classifier while the system was in development, before the results of volunteers' classifications were available. As well, particular scientists have asked the platform developers to run the ML classifier on a set of images on a specific date to check quickly if they can find a correlation between the data and detector instruments. To explain this process, April is identified in this group, said:

> There was something wrong in data and we asked Casper to run ML on yesterday's data and see how many whistles were there and what frequency and time they happened. We were able to do statistics on data and see hundreds of whistles happen at this time and we could look at the instruments at that time and see if there is any correlation.

The final group identified are the **volunteers**, who are affected by ML as its classifications govern what glitches they see. Some of them know what they classify in each workflow has already been classified by the ML classifier. Also, they know that the ML classifier is supposed to learn from volunteers by aggregating their classifications for known classes. However, there are some volunteers who want to know if their contributions really improve the ML classifier. One of them posted on the discussion board:

> Do you have some insight into the effect or lack of effect the human classifications are having on the ongoing machine learning? I'd really like to see more feedback from the LIGO team to help me justify spending my time in this endeavor.

A few of them think that ML classifier should learn the way that they are doing the classification. One of them left a comment on the discussion board:

> I am struggling with this idea a bit... I think that the ML is the one who should learn to adapt to us and not vice versa.

And a few of them tried to learn how the ML classifier would classify an image to make sure that they are classifying correctly and if the ML classifier has something to teach them.

Other perceived uses of the technology In addition to the three implemented functions described above, some of the **volunteers** believed that an ML system has been used to communicate with volunteers on the talk page. Brandon and Katie are identified in this group. Brandon said:

> I read on the Oxford websites ... that they plan in the future to teach autonomous agents who can talk different projects on Zooniverse. Even in Zooniverse, on the main talk forums there are talks about this. So, I think it's happening.

He added that one of the users on Gravity Spy commented on an image in a way that indicates it is a bot. He thinks humans would analyze the image in a different way than what was said about the image.

5.2 Problems with the ML classifier

In this section we describe the problems with the technology as perceived by the members of each of the identified social groups.

ML developers faced different problems designing and implementing the ML classifier. They said that in the early phases of the project they had to retrain the ML model several times on new versions of the gold dataset given by the LIGO scientists, which was computationally expensive. They explained since the ML classifier relies on the gold dataset to learn the classification, it is necessary to retrain it when there is a new gold dataset, but that doing so takes quite a lot computational resources. Later they faced misclassification of some images by the ML classifier caused by some erroneous labels in the gold dataset and an error of the ML classifier.

A first problem of the **platform developers** was to present the results of ML classifier to volunteers or LIGO scientists in an understandable way. Later, they also noticed the misclassification problem of the ML classifier. They said the problem in gold dataset and the ML classifier's algorithm caused the misclassification. But there are also images that could fall in several categories that cause the misclassification. And the current issue is to design a right schema for weighting volunteers' classifications and integrate that to the system to retrain the ML classifier. Making decision on including what parameters is challenging. It affects the score of ML classifier and they need to come up with a framework that has the best impact on the score of the ML classifier. Marsha stated:

> I think that's probably the biggest challenge from the people side is how we can adequately cover all the different parameters that we're able to change and get an understanding of what really affects the results the most.

LIGO scientists were also aware of the primary misclassification of data by the ML classifier and knew the reason was the gold dataset and the ML classifier. They do not have any issues with the current ML classifier and indeed, already use its outputs.

Volunteers also knew about the wrong data in gold dataset that was causing the misclassification by the ML classifier. And there are some volunteers who are concerned if the problem of the ML classifier decreased by training on newcomers' classifications. One of them posted a comment on the discussion board:

> I'd be surprised to learn that GS' problems likely really messed with at least some newbie's classifications. Did those messed up classifications, in turn, mess with the way the ML worked during that time?

They also believe that the ML classifier is not working well for all classes especially in upper level where it does not have a high confidence score and images can fall into several classes.

5.3 Solutions to problems with the ML classifier

Finally, we discuss what members of each of groups perceived as potential solutions to the identified problems.

The **ML developers** improved the algorithm of the ML classifier to handle the problem of the misclassification and trained it over the new gold dataset with corrected labels. This work resolved the problem of misclassification. They know it is expected that the ML classifier classifies all images to the right classes and they used state-of-the-art algorithms to make it more accurate and sufficient. The ML developers think that the ML classifier would have a different result if they add to the training data the glitches that the volunteers have classified. However, they have not yet retrained the ML classifier with the volunteers' data. They believe it is ambitious to not evaluate the results of the ML classifier and trust it without volunteers' evaluations.

The **platform developers** also believe it would be ideal to have the perfect algorithms for ML classifier that are able to classify the images without any needs for evaluations. However, they tried to come up with some solutions to improve the ML classifier in GS. They created a framework to include all volunteers' classifications based on their expertise and assign a credit to each volunteer. They should work on that to see if it improves the result of the ML classifier.

The **LIGO scientists** helped to correct the gold dataset labels, which consequently improve the ML classifier. Since then they are very satisfied with the current results of the ML classifier.

Some **volunteers** approached their problems by understanding how the ML classifier can be improved over known classes. They think they are interacting directly with the ML classifier's result in each workflow and learn what ML is classifying. Brandon said:

> ...I would have classified it in a different category. But I have accepted that the machine classified it that way, and during the learning process, I was trying to learn how the machine thinks. Because sometimes I could be wrong, too; other times, the machine could be wrong. And in each cases, it's especially unclear who is right. Sometimes you just decide.

Regarding images that fall in several classes they said they need a consensus for lots of cases as they should make a decision to have a ground truth for those images. One of them posted on the discussion board:

> If they would fall in either of these category by consensus that would make it easier to the ML to learn and to classify them.

There are other volunteers who try to understand how to improve the ML classifier by proposing some solutions. One of them posted on the discussion board:

> Maybe the machine algorithms could have variables that are a function of weather or time of day or local temperature or magnetic field or whatever may affect the measurement.

6 Discussion

The case presented above has some implications for building such systems and for researching them. First, methodologically, we found that it was useful when documenting the relevant social groups and their perceived problems and solutions to consider what use members of the groups were trying to make of the technology and so their opportunities to learn about it. In the Gravity Spy case, the groups and their relation to the technology are shown graphically in Figure 1. The figure shows that the ML developers are closest to the new technology (the "genie in the bottle"), as they are intimately involved in and try to make it work (coaxing the genie out of the bottle). However, other groups interact with the technology more indirectly. The volunteers, for example, are subject to the decisions of the ML classifier but have no easy way to see how it is designed or how it is performing. As a result, the further away from the bottle, the fuzzier the conception becomes.

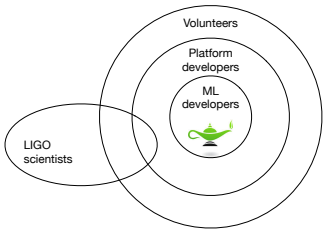

Fig. 1. Circles of engagement with machine learning in Gravity Spy.

Second, and related to the first point, groups with less contact with the technology must rely on other sources of information to make sense of its capabilities. For example, the LIGO scientists do not have the experience of building the ML classifier themselves, as Casper said:

> A lot of people just receive the end-products. There are some inputs. There's a black box and then there's some end-products and they don't think about either the inputs or the black box that led to the end-products. They look all GPS times have labels and people think it's ok. So, taking the output without knowing the input or the black box makes everything blurry.

In short, they see the output, classified glitches, which address a pressing need within their own practice, and not the caveats about performance.

Volunteers have even less opportunity to see how ML is being used as the system does not expose the details of the ML performances to avoid biasing volunteers' own classification. However, this design means that users have no easy way to explore the system's capabilities. Rather, it appears that in making sense of an "ML assistant", they draw on their own experience as contributors to the

project, to scraps of information on various project blogs and to more general publications about AI. A particular confusion seems to be about the difference between narrow and broad AI, one able to do just one task vs. many, leading some to conceive of the ML as filling the role of a participant in the project (i.e., anthropomorphism), not only classifying but also posting and discussing. As a result of this belief, there are interactions in which volunteers believe humans actions are actually those of machines (i.e., bots), what we label "technopomorphism". Given the rapidly advancing capabilities of chatbots, belief in chatbots is not unreasonable, and indeed, there may soon be Zooniverse chatbots, even though there aren't at present. This experience suggests that when the bots do arrive, the identity of the human and machine elements should be made clearly visible to volunteers with labels in spaces where the two interact and tutorials describing where the boundaries of human and machine are, in other words, providing project resources for understanding the genie, even when it is not directly visible.

7 Conclusion

This initial study has examined just one setting with a limited number of interviews. In future work, we hope to expand to more settings and more thorough data collection. As well, our initial findings provide the basis for development of a systematic coding system for the volunteers' posts. Even in its initial state, we believe our study is useful in revealing the difficulties stakeholders in an ML may face in forming an accurate understanding of the system's role and capabilities. Misapprehensions about technology capability are not restricted to Gravity Spy. For example, witnessed by recent crashes, Tesla drivers seem not to universally understand the limits of the Tesla Autopilot (a problem that is not helped by choice of name). These understanding matter because the level of performance that is required or suitable depend heavily on the context. Some error in targeting an ad is okay, in diagnosing a disease less so and in recommending a prison sentence or driving a car, perhaps not at all. But from the outside, a user may not be able to tell how well a system for these different uses is performing. And conversely, the requirements that are apparent to users are less visible to developers, leading to a mismatch between design and expected performance. Future work should consider how to make the limitations of ML more visible to those who interact with its results but not the technology itself. It will be beneficial to have a standardized and easy to understand a way to communicate an ML system's level of performance, something akin to the descriptions of gas mileage found on cars.

References

1. Bahaadini, S., Rohani, N., Coughlin, S., Zevin, M., Kalogera, V., Katsaggelos, A.K.: Deep multi-view models for glitch classification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2931–2935. IEEE (May 2017)

2. Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., Shirk, J.: Citizen science: A developing tool for expanding science knowledge and scientific literacy. BioScience **59**(11), 977–984 (Dec 2009)
3. Boyatzis, R.E.: Transforming qualitative information: Thematic analysis and code development. Sage Publications (1998)
4. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2), 77–101 (2006)
5. Collins, H.M.: The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. Sociology **9**(2), 205–224 (1975). https://doi.org/10.1177/003803857500900202
6. Denzin, N.K., Giardina, M.D.: Introduction. In: Qualitative Inquiry—Past, Present, and Future, pp. 9–38. Routledge (2016)
7. Klein, H.K., Myers, M.D.: A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly pp. 67–93 (1999)
8. Miles, M.B., Huberman, A.M.: Qualitative Data Analysis: An Expanded Sourcebook. Sage Publications (1994)
9. Pinch, T.J., Bijker, W.E.: The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. Social Studies of Science **14**(3), 399–441 (1984)
10. Zevin, M., Coughlin, S., Bahaadini, S., Besler, E., Rohani, N., Allen, S., Cabero, M., Crowston, K., Katsaggelos, A., Larson, S., Lee, T.K., Lintott, C., Littenberg, T., Lundgren, A., Øesterlund, C., Smith, J., Trouille, L., Kalogera, V.: Gravity Spy: Integrating Advanced LIGO detector characterization, machine learning, and citizen science. Classical and Quantum Gravity **34**(6) (2017)