The AAAI 2017 Spring Symposium on
Designing the User Experience of Machine Learning Systems
Technical Report SS-17-04

Not-So-Autonomous, Very Human Decisions in Machine Learning: Questions when Designing for ML

Henriette Cramer, Jennifer Thom
Spotify
{henriette, jennthom}@spotify.com

Abstract

Until the machines are fully autonomous and generate themselves, human design decisions affect Machine Learning outcomes every step of the way. This position paper outlines multiple stages at which design decisions affect machine learning outcomes, and how they interact. This includes: dataset curation and data pipelines, selection of optimization targets, and the designed dialogue with end-users with its implicit and explicit feedback mechanisms. We specifically also call out another user group that appears somewhat overlooked in the research literature – the data curators and editors often involved in selecting and annotating the data that machines learns from.

The human side of machine learning

While Machine Learning-based systems may appear to autonomous learn and make decisions on their own, they are working on people's behalf. Human decisions affect their outcomes at virtually every step of the way. From decisions on data curation, optimization target selection and UX design choices, to end-users' decisions whether or not to take the recommendations that a system provides on board; choices are made. Very few systems are actually fully autonomous and currently human decisions still determine where and how a system is deployed. The ecosystem surrounding any actual machine learning system is a thoroughly human affair, whether consciously *designed* or not. This position paper will outline a number of questions to ask when developing a machine learning-based systems. We won't claim to have the answers, but we do believe it is useful to outline this collection of questions and refer the reader to some literature in each.

How does data curation affect the ultimate end-user experience?

The design of those mechanisms, pipelines and processes to feed the machine with training data help to shape the possible outcomes of a machine learning system. The people who generate this data and their human characteristics affect how well different populations will be served by machine learning systems and which human biases will be reflected by a system's decision making. Should this bias be 'corrected', an open question remains what standards are applied to make these corrections (Snow et al., 2008). In addition, the ideal of neutrality when devising a labeling system does not hold as labels often have systemic values biases coded from within (Bowker and Star, 2000). When labeling for constructs such as "quality", Kay et al. (2015) for example observes that annotators can fall back into heuristics consistent with gender stereotyping. 'Objective' standards may not exist, and may not hold true across a diverse crowd. Designers and design researchers have an opportunity to influence this process by involving themselves in the process of curation and labeling, whether in the design of labeling schemes or in the design of the systems that enable curation and labeling.

Who are the humans inside the machine?

Despite optimistic rhetoric about automation, skilled people do the work of the shaping of data into information that is useful for the modeling and training of the machines. The tools that they use for this work are often primarily designed for high volume data processing and throughput as opposed to other human experience values (Reidsma et al., 2005). Do designers of machine learning systems consider the diverse sets of curators and annotators that possess multiple sources of knowledge? The personal stories of the early curators at Yahoo for example illustrate the impact that their work has had; in particular, the type of local and deep knowledge they had of certain topics as well as the system itself (NY Times, 2016).

Design opportunities exist in this space, if we consider the local knowledge that annotators and curators hold, and leverage this expertise may lead to more diverse systems. For example, designers may also take into account the "voice" of a curator, and their added expertise, when designing the behavior of a machine learning system. By making choices to include or exclude, what type of implicit dialogue occurs between the curator and the ultimate end-user of the system?

How do optimization targets affect the design of machine learning systems?

The design process of machine learning systems also encompasses the choice of metrics used to measure engagement and quality. We, as designers of machine-learning systems, choose to tune algorithms to be successful according to these metrics. There are tradeoffs that are made when one metric is valued over another, and unintended biases may emerge. In addition, high performance on a machine system metric is not often the same as 'success' from an end-user perspective. Optimizing for clicks may lead to clickbait, long-term success and engagement may look different. Yi et al. (2015) for example describe how using dwell time proved a better avenue beyond clicks. Such metrics definition and validation work in finding the metrics that matter in user satisfaction remain an active challenge for especially new types of interaction models. One challenge for designers of machine learning systems and data-driven design is to become conversant in these metrics, such that we can work with those who instrument systems so we can make sure that these systems reflect user behavior in a meaningful way and in turn, give us design feedback to improve the system.

How do implicit and explicit feedback from end-users feed into your optimization targets?

The instrumentation of a system to record implicit feedback, such as behavioral measures, clicks, listens, views, can be fed back into the machine in service of optimization targets. How this information is recorded and employed can affect the end-user experience as what users read, see and hear will be swayed by this feedback. The collection of such implicit feedback for use in experimentation is an industry standard but when made transparent to end-users, can be disconcerting. Designers must also consider explicit feedback, such as stars, thumbs up/down, yelling at the autonomous thing that just won't work, and how best to feed them into optimization targets. Is explicit feedback often a stronger signal of what an end-user wants? How should your agent or system react to strongly expressed notions of end-user preference?

What is made easier or harder to reach?

Depending on the interaction model chosen, machine learning based approaches can make it both easier to reach relevant information, and harder to access any other information that a system deemed less relevant. For example, a recommender model may make a ranking of items, but not all these ranked options may be actually accessible to the end user. This is especially apparent when contrasting a list of blue link search results versus for example a direct answer from a voice assistant. A search or voice assistant giving one direct answer can be great and efficient – but in choosing the succinct answer, there will necessarily be answers that will be hidden. What is difficult to describe or request for your human user in the chosen interaction model? What won't ever surface as your system cannot express it? If you don't gather the training data that allows for the unexpressed answer, then a machine-learning system may never even know that it is missing possibilities that it cannot express. It can however be challenging to even identify if, and which, problems are occurring. Algorithms, outcome presentations and data all have a range of biases (Baeza-Yates, 2016).

Should the black box be made transparent?

If a designer wants to unveil the magic of the machine, there remains many open questions about when, how and if this should be done. Are these explanations technically possible? Should autonomous systems be able to explain themselves in a human-understandable way to the end-user? If the assumption is that these explanations might be helpful to the user, the design features that affect control and understanding should be uncovered and explored. This is non-trivial, and the results of aiming to explain are not always as expected, nor similar to other explanations (Herlocker et al., 2000). Beyond the work on the consequences of different design of explanations – and the work to be able to automatically generate such explanations, if at all possible - current discussions have arisen about for example the EU's directives related to the 'right to explanation' of algorithmic decision making (Goodman & Flaxman, 2016).

Should the machine-learning system adhere to or break social conventions?

Virtual assistants, such as Alexa and Siri, play with the idea that they share some social characteristics with the humans that interact with them. When they don't work as expected or designed, these assistants often play up their machine-like nature and apologize for their perceived shortcomings. For other autonomous systems, these types of social conventions may not always be appropriate. The social cues that emerge in statistical machine-learning systems may often be over-ridden by deliberate rules in order to maintain a certain social character.

What are the implications of these decisions? Social conventions may change if the autonomous system is embodied in a particular way or uses certain types of language. As designers, we have the opportunity to explore play, and messing around, as well as push the boundaries of social conventions. In the context of product, this is a complicated set of discussions between various stakeholders ranging from technical feasibility to brand and market considerations. Just as humans exhibit various personas depending on context and social situations, should we as designers train our machine-learning systems to behave and learn to react in the same ways? Even with the extensive amounts of research in these areas, for which the classic Media Equation (Nass & Brave) and Wired for Speech (Reeves & Nass, 1996) work remain the canonical examples, a lot of questions remain for designers making pragmatic decisions for their specific product and context.

In short.

Humans feed the machine, humans control the machine's desires, and humans consume the products of the machine. And right now, humans still make the decision to turn the machine on or off. There are plenty of design decisions and data choices to consider, and their answers aren't necessarily obvious.

Biography

Henriette Cramer

I'm a sr. research lead at Spotify, where I'm focusing on the human side of machine learning and the dialogue between people, data and machines. My data & design research revolves around people's interaction with systems that learn and adapt, and the resulting feedback loop in both research and applied product settings. I'm particularly interested in the ecosystems that surround individual users' interactions, the effects that different design strategies have on people's perceptions, and the (mis)match between human and machine models of the world around them.

Before joining Spotify, I was part of Yahoo Labs, researching user engagement, mobile & search-related projects. Prior, I was a researcher at the Mobile Life Center in Stockholm, Sweden where I led projects on human-robot interaction and location-based services. My original academic background is in people's responses to adaptive and autonomous systems.

Jennifer Thom

I am a sr. research scientist at Spotify also focusing on the human side of machine learning and in particular, the social and collaborative aspects of the work conducted by those who label and collect the data that underlie these systems. I'm also interested in conversational interfaces and the social aspects of dialogue between humans and machines.

Previously, I was a research scientist at Amazon where I used various crowdsourcing techniques to provide data to improve the machine learning models that power the Alexa assistant and investigated informal question-answer behavior while a research scientist at IBM Research.

References

Baeza-Yates, R. Data and Algorithmic Bias in the Web, IEEE WCCI, Vancouver, Canada, July 2016.

Bowker, G. and Star, S. *Sorting things out*, MIT Press, 2000.

Goel, V. When Yahoo Ruled the Valley: Stories of the Original 'Surfers', July 16, 2016, *New York Times*, http://www.nytimes.com/2016/07/17/technology/when-yahoo-ruled-the-valley-stories-of-the-original-surfers.html.

Goodman, B., Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation", arXiv:1606.08813

Herlocker, J.L., Konstan, J.A., Riedl, J.. 2000. Explaining collaborative filtering recommendations, in *Proceedings of CSCW2000, ACM, New York, NY, USA, 241-250.*

Kay, M., Matsusek, C., and Munson, S. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of CHI2015*. ACM Press.

Nass, C, Brave, S. *Wired for Speech*, MIT Press, 2005

Reeves, B., Nass, C. The Media Equation, 1996. How People Treat Computers, Television, and New Media Like Real People and Places. Univ of Chicago Press.

Reidsma, D., Hofs, D. and Jovanovic, N. "Designing focused and efficient annotation tools," in *Measuring Behaviour, 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, The Netherlands, 2005.

Snow, R., O'Connor, B., Jurafsky, D. and Ng, 2008 A.Y. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254-263). Association for Computational Linguistics.

Yi, X., Hong, L., Zhong, E., Nan Liu, N, and Rajan. S. 2014. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*, ACM, New York, NY, USA, (pp 113-120).