Ways of Knowing for AI

Q. Vera Liao, Michael Muller

IBM Research, Yorktown Heights and Cambridge
vera.liao@ibm.com, michael_muller@us.ibm.com

Abstract: knowledge acquisition is critical for AI models. However, the AI community have largely overlooked the opportunities for interactions between AI models and human knowledge sources, which are currently limited to providing simple labels as oracles. This limited perspective not only restricts the productivity of AI model development, but also hinders the consideration of important issues in the procedures and outcomes of human knowledge acquisition such as data biases and ethics. We argue that for an AI system with the goal of optimizing performance and user experience, it could leverage diverse HCI methods and HCI perspectives to interact with human knowledge sources. We conduct an analysis on a taxonomy of knowledge acquisition problems for AI, and mapping it with HCI methods for acquiring knowledge about a domain or stakeholders. We present a provocative idea of a future AI that embodies itself as an HCI researcher bot to interact with various stakeholders.

Keywords:

- Knowledge acquisition for AI
- Human in the loop
- Machine learning

To put it simply, machine learning (ML) is the process of using standardized learning algorithms to learn from task specific data or knowledge. Despite the enormous effort by the AI research community on advancing the capabilities of learning algorithms, obstacles persist. in the practices of AI application development, the critical blocker is often the acquisition of task specific knowledge, as illustrated in the issues of expensive, insufficient, and noisy labeled data, and the time-consuming processes of data-cleaning and feature engineering.

The HCI community have been responding to these issues since 2000s (Fails and Olsen, 2003). Contrary to the classic AI goal of automation, the alternative solution is an emphasis on human-in-the-loop approaches and a collaborative perspective by viewing "*learning algorithms as interacting with both computational agents and human agents to optimize learning behaviors through these interactions*" (Holzinger, 2016). One example is *interactive machine learning* systems, where a domain expert can make rapid and incremental model updates by, e.g., iteratively supplying small batches of training examples (Amershi et al., 2014). Another example is interfaces supporting *active learning*, where the ML algorithm learns by actively requesting specific information from the humans (Cakmak and Thomaz, 2012).

A recent effort made jointly by some AI and HCI researchers is a call for paradigm shift from machine learning to *machine teaching* (Zhu, 2015; Simar et al., 2017). Instead of treating the humans in the loop as oracles that perform mechanical tasks, they should be seen as teachers who impart task specific knowledge to the learning algorithms. Machine teaching research should thus put an emphasis on interfaces supporting the teachers' interactions with ML models and expression of knowledge, and making the teaching experience more productive and engaging. While the "teacher" role is a valuable change of perspective to start thinking about more meaningful and holistic interactions between humans and ML models, we argue that *the teacher role has limitations*. It assumes that there are dedicated teaching personnel in the model development process. While this may be a proper description of domain experts participating as members of AI development teams, it may be less applicable to other, more transient human roles that ML models can acquire knowledge from, such as crowd workers, end users and the many people who may be involved in the task but not available to work directly with the model. Moreover, "teaching" implies teacher-initiated interactions and puts the ML model in a passive "learner" role, and presumes knowing of teaching goals and learner status. This, again, may be less applicable to the other categories of human-in-the-loop roles.

In this extended abstract, we start developing requirements for a general interface between ML models and diverse human knowledge sources, and envision *how an AI that aims at continuous self-learning for improvement of system performance and user experience should embody itself to interact with targeted users, domain experts and diverse stakeholders*. This vision is motivated by the prediction of three upcoming trends of AI:

- First, as the learning capabilities of AI algorithms evolve, AI should learn from **rich forms of human knowledge** beyond the current practices of instance labeling (Ratner et al., 2016), which is highly constrained and inefficient, and fails to make use of more tacit knowledge such as heuristics, rationales and common sense (Polanyi, 1962).
- Second, as the model construction and optimization become increasingly automated (Biem et al., 2015), AI will take **self-initiative** to acquire human knowledge instead of relying on model developers. Meanwhile, to avoid catastrophic risks, the autonomous AI should actively seek control from and co-operate with humans. So the human knowledge is not only a source of learning materials, but also for evaluation, validation and for the AI to build common ground with humans, and thus should be sought carefully and frequently.
- Lastly, AI applications will become increasingly versatile and composeable. For example, an autonomous car would require training interconnected ML models for visual, reasoning and motor skills. It could benefit from having a **common interface** to interact with different human knowledge sources for different ML problems.

We argue the premise for such an interface is a *toolbox of knowledge elicitation methods* with the goal of learning the task domain, understanding diverse stakeholders, and ultimately, optimization centered around user needs and societal benefits. HCI is a discipline that predominately focuses on developing such a toolbox (Olsen and Kellogg, 2014). By considering the types of knowledge that current or near-future ML models need to acquire from human knowledge sources, ranging from explicit to tacit, and the types of human roles involved, we propose a taxonomy of knowledge elicitation for ML models. We then attempt to map HCI methods to the same taxonomy (Figure 1). This gives us a starting point to consider interactions for acquiring human knowledge for different kinds of ML problems.

With that, we invite the idea of a future AI that *embodies itself as an HCI researcher bot*, and that selectively uses HCI methods to acquire system knowledge from its targeted users, domain experts and other stakeholders. We think this AI entity *not* as a replacement for an HCI expert, but rather as an embodied extension of the human HCI experts. However, we would like to leave out the discussions on the system operation and its relation with human HCI researchers, i.e. whether it is an autonomous system, a delegation, or a hybrid co-investigator, but focus on considering its potentials as an *interface* that can potentially satisfy the above requirements---eliciting rich forms of knowledge, taking self-initiative, and sharing a common form supporting different ML processes. This interface could help us rethink the interactions between AI models and human knowledge sources. Besides harnessing the numerous lessons from decades of HCI research on "ways of knowing" (Olsen and Kellogg, 2014), it redefines the relationship between AI models and human knowledge sources not as algorithms and oracles, nor students and teachers, but investigators (whose ultimate goal is to optimize the system) and informants. Such an embodiment also encourages the adoption of HCI best practices on how to interact with human subjects, including the design of "probes" (Boehner et al., 2007) and study procedures, ethics considerations, and sampling methods.

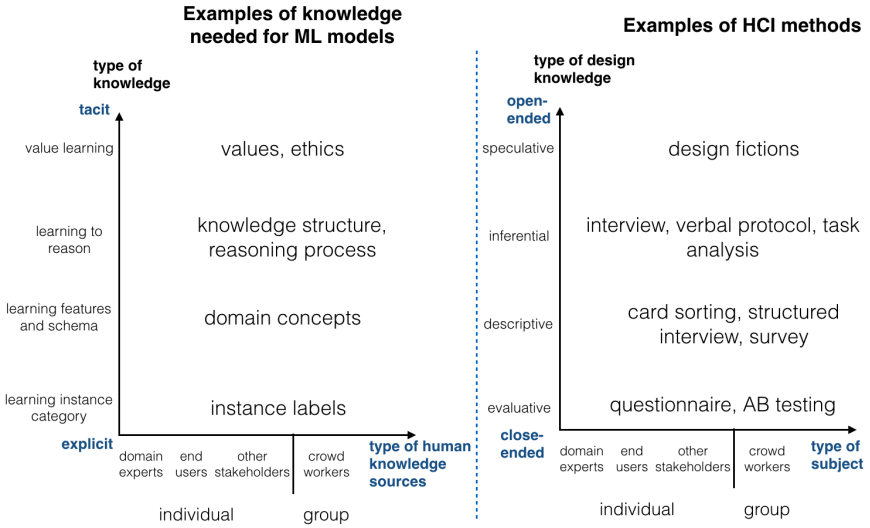

Figure 1. A taxonomy of knowledge elicitation for ML models and mapping of HCI methods

In the following, we briefly discuss four cases of ML knowledge acquisition problems in different places of the taxonomy, and envision possible solutions provided by the HCI researcher bot.

Learning instance labels

This is currently the most common ML tasks as represented by supervised learning algorithms that infer a predictive function from a set of labeled instances. This maps directly to experimental and survey methods where the measurement (labels) is explicitly defined (i.e., as "ground truth") and data points (training data) are collected for quantitative analysis (modeling). In fact, data scientists frequently collect labeled data by designing experiments or questionnaires (broadly, including simple questions for label selection).

To acquire the knowledge about instance labels, the bot can replace the current data collection efforts of data scientists, by performing experiment design and execution, and administration of surveys. Recent critical works on data scientists' data collection practices, including its situated and contextualized nature as "design work" (Feinberg, 2017) and the pervasive problem of uncertainty (Boukhelifa, 2017) elucidate some challenges in the bot's work.

Schema and feature learning

The ML problem in this category is to learn descriptive knowledge of a task domain, such as a taxonomy or action space, or the feature space. For example, in developing a virtual assistant AI system, the first task is to define the schema of tasks that users may need assistance with. This is currently done by either expert input, clustering or pattern extraction algorithms, or a combination of both (e.g. Hoque and Carenini, 2015).

Defining domain schema and mapping action space are familiar tasks for HCI researchers. The bot can perform observational studies, task analysis and contextual inquiry with targeted users. The bot can also request to access behavioral data from existing systems or tasks. It is noteworthy that there is a long tradition of research on analytical methods for task analysis and inducing concept schema from qualitative data (Rossen and Carroll, 2002; Adams et al., 2008), and there is an ongoing discussion on the complementary offering of these approaches (Baumer et al., 2016; Muller et al., 2016). These discussions may inform how the bot should construct its data collection protocols and make situated interpretation of the task domain representation.

Learning to reason

Learning to perform complex reasoning tasks is in the current frontier of AI research to achieve higher level of intelligence. It requires formalizing the often-tacit knowledge of human reasoning processes. The formalization needs to be performed in at least two areas. One is the construction of detailed domain knowledge, such as a knowledge graph. The other is to learn the inference procedures. While human-in-the-loop systems of this kind are still rare, two relevant areas explored human knowledge acquisition methods: 1) One is expert systems that proliferated in 1980s that aim to emulate reasoning processes of human experts with static rules. To understand how experts make decisions, a combination of interviews, focus group, scenario exploration and observations are adopted (Cooke 1994). These methods are often used for formative research in HCI, which aims to understand the latent user needs, behaviors and motivation. 2) Another relevant area that AI researchers started to explore is crowdsourcing knowledge representations (e.g., Witbrock et al., 2013). This has proved to be a challenging task to achieve accuracy, coverage and efficiency without carefully designed questions and iterated data-collection with quality checks. A critical and under-explored issue is biases in the knowledge representation following bias-insensitive data collection procedures, for which we expect some of the HCI traditions, such as feminist HCI (Bardzell and Bardzell, 2011) and Indigenous methodologies (Kovach, 2009), would be able to provide solutions.

We envision the capabilities to learn from human knowledge sources to perform reasoning tasks as a critical requirement for future (general) AIs. This opens up a challenging design space for knowledge acquisition interfaces of ML models, as the targeted knowledge moves to the tacit end of the knowledge spectrum, and the data space becomes more open-ended. It requires the bot to be versatile in using diverse research methods and adept in designing study protocols and effective "probes".

Value learning

On the very tacit end of the knowledge spectrum, we consider the emerging AI topic of the "value alignment problem" and the following needs for value learning (Greene et al., 2016; Hadfield-Menell et al., 2016). There is a growing interest in both the scientific community and the public to ensure the goals of autonomous AI systems to be aligned with human values (Russell et al., 2017). Given the complexities of human value systems, researchers started to explore learning tacit values from human actions. Because it is extremely hard to obtain large amounts of human choice data bounded by value structures, some researchers proposed novel data collection methods to "crowdsource" choices from speculation of fictional scenarios. For instance, the “trolley problem” is a widely used scenario of ethical dilemma, for which large crowdsourced datasets have been created in the hope of informing the development of “moral machines” (Bonnefon et al. 2016; Shariff et al. 2017).

In line with these ideas, speculative methods in HCI has a tradition of focusing on inquiries about values associated with new technologies. Recently, Muller and Liao (2016) proposed using design fictions as probes, participatory construction and group co-creation to inquire about value issues around AI technologies (for related approaches, see Sorell and Draper, 2014; Cheon and Su, 2016). The bot can use these tools, for example, by generating contextualized design fictions that are formatted for value elicitation (e.g., strategically incomplete decisions, contrasting values).

Reference:

Adams, A., Lunt, P., Cairns, P. (2008). A qualitative approach to HCI Research. In Paul Cairns and Anna L. Cox (eds.), *Research methods for human-computer interaction*. Cambridge University Press.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4), 105-120.

Bardzell, S., Bardzell, J. (2011). Towards a feminist HCI methodology: Social science, feminism, and HCI. *Proc. CHI 201*1, 675-684.

Baumer, E.P.S., Mimno, D., Guha, S., Quan, E., Gay, G. (2016). Comparing grounded theory and topicmodeling: Extreme divergence or unlikely convergence? *JASIST* 68(6), 1397-1410.

Biem, A., Butrico, M., Feblowitz, M., Klinger, T., Malitsky, Y., Ng, K., ... & Sow, D. M. (2015, January). Towards Cognitive Automation of Data Science. *In AAAI* (pp. 4268-4269).

Boehner, K., Vertesi, J., Sengers, P., Dourish P. (2007). How HCI interprets the probes. *Proc. CHI 2007*, 1077-1086.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573-1576.

Boukhelifa, N., Perrin, M. E., Huron, S., & Eagan, J. (2017). How data workers cope with uncertainty: A task characterisation study. *Proc CHI 2017*, 3645-3656.

Cakmak, M., & Thomaz, A. L. (2012). Designing robot learners that ask good questions. *Proc. HRI 2012*. 17-24.

Cheon E., Su N. M. (2016). Integrating roboticist values into a value sensitive design framework for humanoid robots. *Proc. HRI 2016*, 375-382.

Cooke N.J. (1994) Varieties of knowledge elicitation techniques. Int. *J. Human-Computer Studies*. 41. 801-849.

Crawford K. and Cato R. (2016). There is a blind spot in AI research. *Nature* 538 (7625).

Fails, J. A., & Olsen Jr, D. R. (2003). Interactive machine learning. *Proc. IUI 2003*. 39-45.

Feinberg, M. (2017). A Design Perspective on Data. Proc. CHI 2017. 2952-2963.

Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., Williams, B. (2016). Embedding ethical principles in collective decision support systems. Proc. AAAI 2016, 4147-4151.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Proc. NIPS 2016*. 3909-3917.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. *Brain Informatics*, 3(2), 119-131.

Hoque, E., Carenini, G., (2015). ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations, *Proc. IUI 2015*. 169–180.

Kovach M. (2009). Indigenous methodologies: Characteristics, conversations, and contexts. University of Toronto Press.

Muller M., Guha S., Baumer, E.P.S., Mimno D., Sharmi, S. (2016). Machine learning and grounded theory method: Convergence, divergence, and combination. *Proc. GROUP 2016*, 3-8.

Muller M., Liao Q.V. (2017). Exploring AI values and ethics through participatory design fictions. Presentation at HCIC 2017.

Olson, J. S., & Kellogg, W. A. (Eds.). (2014). *Ways of Knowing in HCI* (Vol. 2). New York, NY, USA:: Springer.

Polanyi M. (1962). The tacit dimension. Doubleday.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., & Ré, C. (2016). Data programming: Creating large training sets, quickly. *In NIPS 2016*. 3567-3575.

Rosson M. B., Carroll J. M. (2002). Usability engineering: Scenario-based development of human-computer interaction. Academic Press.

Russell, S., Dietterich, T, Horvitz, E., Selman, B., Rossi, F., Hassabis, D., Legg, S., Suleyman, M., George, D., and Phoenix, S. 2017. Open letter to the editor: Research priorities for rubust and beneficial artificial intelligence: An open letter. *AI Magazine.* 36(4),

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, 1.

Simard, P. Y., Amershi, S., Chickering, D. M., Pelton, A. E., Ghorashi, S., Meek, C., ... & Wernsing, J. (2017). Machine Teaching: A New Paradigm for Building Machine Learning Systems. arXiv preprint arXiv:1707.06742.

Sorell T., Draper H.,(2014). Robot carers, ethics, and older people. *Ethics and Information Technology*, 16(3). 183-195,

Suchman, L. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge University Press.

Witbrock, M., Baxter, D., Curtis, J., Schneider, D., Kahlert, R., Miraglia, P., ... & Vizedom, A. An interactive dialogue system for knowledge acquisition in cyc. *Proc. IJCAI 2013*

Zhu, X. (2015). Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education. *In AAAI 2015*. 4083-4087.