Human-AI Collaboration: Towards Socially-Guided Machine Learning

Q. Vera Liao
IBM Research AI
vera.liao@ibm.com

Rachel K.E. Bellamy
IBM Research AI
rachel@us.ibm.com

Michael Muller
IBM Research AI
michael_muller@us.ibm.com

Heloisa Candello
IBM Research
hcandello@br.ibm.com

ABSTRACT

As Machine Learning (ML) systems become increasingly ubiquitous, capable and autonomous, it has become essential to take a human-centered view to consider how people's interactions with ML systems, including the effort to develop and evolve ML systems, impact their work practices, wellbeing and the social-organizational environment. Built on our work on human-agent collaboration, we suggest a change of perspective, by considering human(s) and the ML model(s) they interact with as a team engaging in collaborative work. With that, we can apply metaphoric thinking based on team collaboration to inform the design of human-model interactions and rethink the collective goals to be embedded in computational models. Based on pillars of Computer-Supported Cooperative Work (CSCW) research, we point to three areas for future research into technologies to support human-model interaction as collaborative work: (1) model training as knowledge sharing; (2) interactions as communication actions; and (3) coordination for better collaboration and construction of trust.

INTRODUCTION

Human interactions with Machine Learning systems are growing to be ubiquitous as driven by two trends. One is the current movement for "democratizing ML" by lowering the barrier of entry to ML

KEYWORDS

Machine Learning; artificial intelligence; human in the loop; agents; human-AI collaboration.

so that model development will no longer require extensive training nor mastering a specialized programming language. Another is the rapid advance of intelligent technologies that continuously learn from activities or direct training provided by end users. In these current or near-future systems radical changes in interaction design are likely, both to make the process more accessible, and to address fundamental issues in data, learning, evaluation, etc. to make people more receptive to and trusting of ML systems. These radical changes require changes of perspectives. For example, a recent call for a paradigm shift from "machine learning" to "machine teaching" [17] is prompting the research community to focus on supporting the efficacy and wellbeing of machine teachers, i.e., people who build ML models. This point of view has inspired both systems that are more user/teacher friendly, and algorithms that are more labor efficient to train.

In this essay, we propose another change in perspective, by considering human(s) and the ML model(s) they interact with as a team, where members have different expertise and strengths; and model development, in the broad sense of learning and evolving, is treated as collaborative work for accomplishing the targeted task. For example, while humans have rich domain knowledge and better understanding of the social and cultural contexts the system is embedded in, a ML model promises greater discipline by making data-driven decisions [7]. Team work then becomes a *metaphor* to guide user interface design for human-model interactions [3]. Metaphor is a useful HCI concept to give users instantaneous knowledge to interact with an unfamiliar system [4]. We are likely to find more parallels with such a metaphor as ML systems become more autonomous. More importantly, this collaborative perspective could help us rethink both the goals of ML development and various aspects of interactions by drawing inspirations from the large volume of literature on human collaboration and CSCW. Specifically, it could help:

- Consider collective goals in intelligence, task performance and human wellbeing, to redefine optimization functions and metrics.
- Develop interaction techniques based on activities in human collaborative work, which could also inspire new computational models.
- Foresee potential issues and borrow design guidelines from theories and best practices in human collaboration and technologies to support it.

The human-agent interaction community (including HRI) have long been following these approaches to closely model human interactions both in terms of system actions and goals [5], including our own work in the past few years [1][3][9][16][15][20]. Some of the work is naturally motivated by anthropomorphic inclinations, where agents are essentially personified interfaces for the underlying ML models. Meanwhile, research on "computers as social actors" (CASA) shows that personification is not a prerequisite for users to apply human social rules in interactions [14]. While some agents are merely using the ML models, the growing areas of adaptive agents and teachable agents are directly concerned with interactions to train or develop the ML models. In this essay, we draw from our experience designing human-agent collaboration and teachable agents to consider lessons for broader areas of human interactions to develop ML systems.

Table 1: McGrath's topology of group modes and functions

	Production	**Group well-being**	**Member support**
Inception	Production demand and opportunity	Interaction opportunity	Inclusion opportunity
Problem solving	Technical problem solving	Role definition	Position and status
Conflict resolution	Policy resolution	Power distribution	Contribution distribution
Execution	Performance	Interaction	Participation

Team collaboration: rethinking goals and activities of ML model development

In designing agents that perform social roles, one may start with considering models that describe activities of humans performing the same roles. For example, in our work to design conversational agents as team members in group decision making (e.g., a facilitator), we have found McGrath's model characterizing team behaviors [13] an invaluable framework to both inform the design of the agent's actions in the team, and to define its high-level goals, including the computational functions.

We could also see similarities between these group modes and the typical process of ML model development. Inception may correspond to the initial problem formulation stage where one explores the production or task demand and opportunities to serve the demand, such as what kind of data can be acquired. The problem-solving stage may correspond to the major chunk of modeling work such as data cleaning and featuring, while conflict resolution can be seen as the debugging work where the model developers (and subject matter experts) resolve inconsistencies between the model output and their knowledge about the task. Finally, the execution is about team performance, i.e. how well the resulted ML systems complete the targeted task.

While this mapping supports the parallel between model development and team collaboration, and the metaphoric thinking for interaction designs, it is important to understand that the key utility of McGrath's model is to shift the attention from the lower-left cell (obsession with performance) to other rows and columns that may have consequences for collaboration. While there may not be a precise mapping, it is crucial for human centered machine learning to pay attention to the other two columns. In this context, to support group well-being means to *support the desired interactions and ecosystem between human(s) and the ML system(s)*, which may involve intricate issues such as role definitions and affordance of interactions. To focus on member support means to *direct more attention to individual user and ML model's participation*, which requires careful consideration for human control on autonomous ML systems, as well as coordination between different human roles involved in the ML development process.

This collaborative perspective also opens doors for drawing inspiration from technologies that support team collaboration, which are primarily studied in the academic discipline of CSCW. CSCW research is often considered to have three pillars: knowledge sharing, communication and coordination. Accordingly, in the following sections, we discuss three foci for future research to support human-model interaction as collaborative work: *model training as knowledge sharing*, *interactions as communication actions*, and *coordination for better collaboration and construction of trust*. We will discuss some useful concepts from CSCW research and demonstrate how they can inform the design of interactions with agents and ML systems in general.

Model training as knowledge sharing

A critical aspect of teamwork is to share knowledge that individual members have and to harness the collective knowledge to solve the problem. CSCW work focused on supporting knowledge externalization in the form of technology artifacts or information repositories. Such a knowledge

centered view led to numerous technological tools that aim to *scaffold* the externalization of individual knowledge of different characteristics (e.g.,[19]).

Knowledge is at the very core of ML systems and developing ML models is essentially a process of knowledge sharing from human knowledge sources. The knowledge that a ML system learns from is commonly thought of as the data *labels* provided for supervised learning. However, knowledge is being shared from human(s) throughout the whole development process. The defining of concept (class) and schema, selection of feature and samples all represent knowledge sharing from the people involved. Also, the type of knowledge being shared is likely to become richer as new human-in-the-loop ML models are developed, especially in the areas of interactive machine learning and machine teaching. For example, in our recent work on new tools to train conversational agents, we employed weak supervision algorithms that allow a human trainer to provide high-level *rationales* instead of a large quantity of *labeled data* to build the classifiers for the agent's language understanding capability. While the tool can significantly reduce the time and human cost relative to performance compared to the conventional instance labeling approach, new issues emerged both regarding model performance (e.g., robustness across different trainers) and user needs [11].

We encourage research on designs for knowledge centered interactions with ML systems. This is likely a rich design space given that for a particular type of knowledge (e.g., instance, rules, schema) there may be multiple elicitation methods (e.g., explicit definition, demonstration, feedback, critique, etc.), each of which may require a set of design guidelines. Another critical lesson to learn from CSCW research on knowledge sharing is to avoid a techno-centric view but pay attention to the social contexts, such as the context where knowledge externalization happens and the politics it may bring, as well as the maintenance and update of knowledge for long-term use (See [1] for an overview)

Interactions as communication actions

Communication is the direct interactions between individuals in a team, often through text-based or spoken conversations, and can be either synchronous or asynchronous, co-located or distant. In viewing human-model interaction as teamwork, we draw on work that has applied concepts and patterns in human communication/conversations to the design of interactions. While we are not the first to suggest a conversational approach to human-computer interaction [10], it may play a more valuable role in ML systems given their potential intelligence, autonomy and initiative-taking. Social science research provides rich accounts of human conversational patterns, and they can be applied to designing both specific system actions and general rules, policies or goals for the computational models. For example, one of the basic principles that govern human conversations is common ground [6], which views conversations as to collectively achieve mutual knowledge. Speakers constantly assess if there is clear enough mutual understanding by evidence (e.g., relevant next turn or explicit acknowledgement), and if not, a grounding process (i.e., repair) will be initiated. In human-model interaction, the communication partner is a ML system, and there is a significant mismatch between the ML model and the human's knowledge/mental model of how the system learns, and often the ML model is a "black

box", so completely hidden from the human. This could be seen as the fundamental challenge in designing effective interactions for developing or debugging ML systems.

In our recent work in designing repair strategies for breakdowns in human-agent interactions [2], we suggest the key lies in shifting from one-way interaction, where the user blindly attempts to repair, to two-way interactions where the agent system should also take the initiative to contribute to the repair process. Based on the theory of common ground, we suggest three levels of contributions from the system: 1) explicitly signaling the breakdown, 2) providing resources to assist user repair (e.g., explaining the current model), and 3) proactively suggesting ways to repair. We demonstrate increasing user satisfaction with higher level (3) of contribution from the system.

Similar guidelines can be applied to developing ML systems in general. Instead of relying on one-way human intervention or debugging, it would be beneficial to provide system-initiated mechanisms to support *monitoring*—to identify model issues in a detailed and timely fashion, *transparency*—to help people understand the current model, e.g., through explanation, and *bridging*—to proactively help connect users' mental model and the system model.

Coordination for better collaboration and construction of trust

Coordination is concerned with mechanisms for individuals to conduct interdependent work in cooperative activities, "the act of working together harmoniously" [12], or "the work needed to allow the work to be done"[18]. We can consider issues around autonomy and human control as coordination work. It is especially pressing for auto-ML systems, to consider how human(s) and model(s) should structure and coordinate their work, make appropriate delegation, etc. Establishing productive and satisfactory coordination is not only necessary for good team performance, but also critical for group wellness and building long-term trust among team members. Similarly, in the context of model development, even if much of the ML work can be automated, people's sense of agency, control and close involvement in the process is necessary to establish trust in the final model.

An important lesson from CSCW work on technologies supporting coordination is that flexibility is necessary to deal with the unpredictability that often emerges in real world contexts [18]. Technologies that are built on rigid coordination mechanisms often do not work outside the lab. We suggest similar consideration in tools that support human-model collaboration, such tools should avoid rigid mechanisms and instead provide rich affordance for humans to monitor and intervene at different points of the model development process. Meanwhile, it may be worthwhile to consider mechanisms used by technologies to help regulate collaborative behaviors and establish trust. One such mechanism is elucidated by Social Translucence Theory [8]. The theory outlines three principles that we can potentially borrow for a ML system to establish trust from the people it interacts with: *visibility*—to make current status available; *awareness*—to act according to expected social rules based on cues from the human partner; and *accountability*—to clearly identify its action scope, responsibility and attribution of problems.

CONCLUSION

To conclude, we suggest considering human(s) and the ML model(s) they interact with as a team engaging in collaborative work. Such a perspective highlights the design metaphor of collaboration and looks to existing work in the fields of collaboration and CSCW as theory and design inspiration.

REFERENCES

[1] Mark S Ackerman., Juri Dachtera, Volkmar Pipek, and Volker Wulf. "Sharing knowledge and expertise: The CSCW view of knowledge management." *Computer Supported Cooperative Work (CSCW)* 22, no. 4-6 (2013): 531-573.

[2] Zahra Ashktorab, Mohit Jain, Q Vera Liao, Justin D. Weisz. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdown. *In Proceedings of CHI 2019*.

[3] Rachel KE, Bellamy, Sean Andrist, Timothy Bickmore, Elizabeth F. Churchill, and Thomas Erickson. Human-Agent Collaboration: Can an Agent be a Partner?. *In Proceedings of CHI 2017*, 1289-1294. ACM

[4] John M Carroll, Robert L. Mack, and Wendy A. Kellogg. Interface metaphors and user interface design. *In Handbook of human-computer interaction*, pp. 67-85. 1988.

[5] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost, eds. Embodied conversational agents. MIT press, 2000.

[6] Herbert H Clark, and Susan E. Brennan. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991): 127-149.

[7] Robyn M Dawes, David Faust, and Paul E. Meehl. Clinical versus actuarial judgment. *Science* 243, no. 4899 (1989): 1668-1674.

[8] Thomas Erickson, and Wendy A. Kellogg. "Social translucence: an approach to designing systems that support social processes." ACM transactions on computer-human interaction (TOCHI) 7, no. 1 (2000): 59-83.

[9] Q Vera Liao., Muhammed Masud Hussain, Praveen Chandar, Matthew Davis, Marco Crasso, Dakuo Wang, Michael Muller, Sadat N. Shami, and Werner Geyer. All Work and no Play? Conversations with a Question-and-Answer Chatbot in the Wild. *In Proceedings of CHI 2018*. ACM, New York, NY, USA, vol. 13. 2018.

[10] Paul Luff, David Frohlich, and Nigel G. Gilbert, eds. Computers and conversation. Elsevier, 2014.

[11] Neil Mallinar, Abhishek Shah, Rajendra Ugrani, Ayush Gupta, Manikandan Gurusankar, Tin Kam Ho, Q. Vera Liao et al. Bootstrapping Conversational Agents With Weak Supervision. *In Proceedings of IAAI 2019*.

[12] Thomas W Malone, and Kevin Crowston. "What is coordination theory and how can it help design cooperative work systems?." In Proceedings of the 1990 ACM conference on Computer-supported cooperative work, pp. 357-370. ACM.

[13] Joseph E McGrath. "Time, interaction, and performance (TIP) A Theory of Groups." Small group research 22, no. 2 (1991): 147-174.

[14] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. Computers are social actors. *In Proceedings of CHI 1994,* 72-78.

[15] Claudio S Pinhanez., Heloisa Candello, Mauro C. Pichiliani, Marisa Vasconcelos, Melina Guerra, Maíra G. de Bayser, and Paulo Cavalin. "Different but Equal: Comparing User Collaboration with Digital Personal Assistants vs. Teams of Expert Agents." arXiv preprint arXiv:1808.08157 (2018).

[16] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. Face Value?. Exploring the Effects of Embodiment for a Group Facilitation Agent. *In Proceedings of the CHI 2018*, 391. ACM,

[17] Patrice Y Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos et al. "Machine teaching: A new paradigm for building machine learning systems." arXiv preprint arXiv:1707.06742 (2017).

[18] Anselm L Strauss. "Work and the division of labor." In Creating Sociological Awareness, pp. 85-110. Routledge, 2018.

[19] Christian Wagner,. "Wiki: A technology for conversational knowledge management and group collaboration." *Communications of the association for information systems* 13, no. 1 (2004): 19.

[20] Yunfeng Zhang , Q. Vera Liao, and Biplav Srivastava. "Towards an Optimal Dialog Strategy for Information Retrieval Using Both Open-and Close-ended Questions." *In Proceedings of IUI 2018*. 365-369. ACM