Reciprocal Human Machine Learning (RHML): Human-AI Collaboration Based on Theories of Dyadic Learning

David G. Schwartz [1], **Dov Teéni** [2], **Inbal Yahav** [2]
[1] Bar-Ilan University, Israel
[2] Tel-Aviv University, Israel
david.schwartz@biu.ac.il, teeni@tauex.tau.ac.il, inbalyahav@tauex.tau.ac.il

Abstract

In this position paper we advocate a Reciprocal Human Machine Learning paradigm based on two theories of human-human learning behavior. Drawing from Jörg's theory of reciprocal learning in dyads and the Jewish tradition of Havruta - pair-based study, we suggest that human-machine collaboration based on these established human-human collaborative forms can achieve a rich and robust human-in-the-learning-loop (HITLL) framework in which both parties experience learning over time.

Introduction

Throughout human history, learning at its most abstract level has primarily followed one of three approaches. *Instruction*, be it from parent to child as in the biblical imperative "thou shalt teach them diligently to thine children" (Deuteronomy 6:7), or from teacher to student. This approach, akin to supervised machine learning, involves one who teaches and an other who learns. *Self-directed learning*, in which a person teaches oneself based on the environment, written materials, or other existing and accessible forms of knowledge. This approach, akin to unsupervised machine learning, involves an individual who acts as both teacher and student. *Peer-based learning* in which dyads or small groups seek to gain understanding and knowledge through mutual analysis, discussion, and debate. This approach, which we believe to be a promising model for human-in-the loop (HITL) computing, may lead to intelligent systems in which not only does the machine learn from its human creators/trainers, not only do human users learn from the machine model outputs, but a form of mutual or reciprocal learning can emerge in which each learns from the other in multiple cycles over time. It is this third approach, in which we treat humans and machines as peers engaged in mutual learning, that we consider in our research.

Reciprocal Human Machine Learning (RHML) is a novel approach to interactive learning that builds on Jörg's theory of reciprocal learning in dyads (Jörg 2004, 2009). RHML is also inspired by the Jewish tradition of Havruta, a form of collaborative learning that involves studying a text with a partner and engaging in critical questioning and argumentation (Kent 2010; Holzer and Kent 2013). Both Jörg's theory and the Havruta tradition are firmly rooted in human-human collaboration which, through the RHML paradigm, are brought to bear on human-machine collaboration. RHML aims to create a symbiotic relationship between humans and machines, where both agents learn from each other and co-construct knowledge through feedback and dialogue.

In this position paper, we argue that RHML can address and advance a key aim of the Building Connections symposium: fostering human-machine collaboration. We outline the main components and challenges of RHML and provide some preliminary examples of how RHML can be implemented in practice. We invite researchers and practitioners from different disciplines to join us in exploring the potential of RHML for NLP and beyond, for example image processing.

Why

Why is it important that humans engage in a learning process as part of human-machine interaction? This goes to some of the fundamental problems and biases of machine learning. When left to function alone over time, the accuracy ML models are known to deteriorate as the environment around them changes. HITL advocates and expects periodic human intervention to update or retrain a ML model. We argue that the human collaborator or expert will do this more effectively when they have (a) learned previously unknown insights into the data from the machine; (b) challenged the machine when receiving results they consider incorrect; and (c) are motivated by improvements in their own understanding and expertise. We seek HITLL - a human in the *learning loop*.

Reciprocal Human Machine Learning

An RHML configuration (Figure 1) consists of four main components: (1) a human learner (HL), who provides natural language inputs and outputs to the machine; (2) a machine learner (ML), who processes the inputs and outputs using NLP techniques and algorithms; (3) a feedback mechanism, which allows the human and the machine to exchange information (dialog) and evaluate each other's performance, and (4) Joint representations of classification knowledge that en-

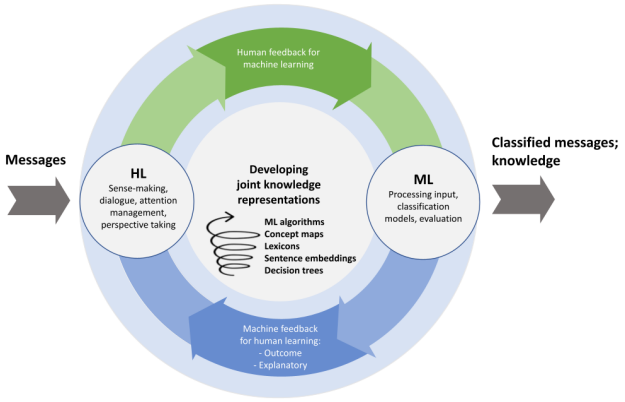

Figure 1: a model of Reciprocal Human-Machine Learning

ables the representation, processing and exchange of information, which is essential for human and machine learning.

The main challenges of RHML are (a) to achieve an efficient and effective balance between human and machine agency, such that both agents can benefit from each other's strengths and compensate for each other's weaknesses; and (b) to support and foster joint knowledge representations that can be interpreted by both machine and human. For example, humans are good at understanding context, pragmatics, and common sense, while machines are good at handling large-scale data, complex computations, and statistical patterns. However, humans may also have biases, errors, or gaps in their knowledge, while machines may also have limitations, uncertainties, or ambiguities in their outputs. Therefore, RHML requires a careful design of the feedback and dialogue mechanisms, which should be adaptive, transparent, and trustworthy. Moreover, such a balance may need to be adapted dynamically according to the relevant circumstances.

To illustrate how RHML can work in practice, we present three scenarios of RHML applications in NLP, two of which are at different stages of implementation and testing, and one of which is illustrative. The illustrative scenario: (1) a human-machine poetry generation system, where the human provides a topic or a theme and the machine generates a poem. The scenarios being actively pursued and tested: (2) text message classification into suspect/non-suspect categories for a cybersecurity application; and (3) in-context sentence annotation in which human and machine conduct collaborative writing. In each scenario, RHML supports the human and the machine with feedback and dialogue to improve their outputs and learn from each other.

Developing and testing for the RHML Paradigm

Our work involves creating human-computer environments that can support the characteristics of the RHML paradigm, and experimental testing of efficacy through multiple learning cycles in which human and machine collaborate to improve classification models. To date we have implemented one RHML environment, Fusion, in which we are testing the approach on a series of different text classification problems (Te'eni et al. 2023,forthcoming). The design of Fusion itself was reported in (Zagalsky et al. 2021) and we refer the reader to that work for a detailed description. The devil is in the details, for it is one thing to conceptualize some form of collaboration, and quite another to implement this on a software platform through which human and machine can effectively communicate, learn and guide change.

Background: Human-to-Human Learning

One of the challenges of human-AI collaboration is how to design systems that can support and enhance *human learning*. Human learning is often an implicit result of human-human collaboration, and we ask ourselves if this should not also be the case in human-machine collaboration - perhaps explicitly so. In this context, it may be useful to consider some well-known examples of human-human learning, such as Jörg's theory of dyadic learning and Havruta. These are two forms of collaborative learning that involve pairs of learners interacting and co-constructing knowledge through dialogue and feedback.

Jörg's Theory of Dyadic Learning

Jörg theory of dyadic learning (Jörg 2004) proposes that learning occurs when two individuals engage in a mutual exchange of perspectives and experiences and use each other as resources for understanding and problem-solving. Dyadic learning can also provide an autonomous learning context, in which learners regulate their own learning behavior through interactions with their peers (van de Sande et al., 2018). Jörg argues that dyadic learning fosters cognitive and social development, as well as metacognitive and self-regulatory skills.

According to Jörg's theory, reciprocal learning is a process of co-creation that entails 'a complex of self-generative, self-sustaining processes of mutual "bootstrapping" with potentially nonlinear effects over time' (Jörg 2004, 2009) Jörg's theory of learning in human dyads models Vygotsky's interactive learning theory as what Jörg calls 'a set of general laws of reciprocal causal interaction'. Vygotsky claimed that the development of human intelligence is achieved by interactively learning from others and co-producing an understanding of the world, rather than by individually accumulating separate pieces of knowledge (Vygotsky 1978). Jörg's laws reflect reciprocal influences that happen during real-time interactive learning held in an educational context. It is based on the social principle of reciprocity and on the cognitive mechanisms underlying complex human psychology. Jörg envisages an architecture in which two identical intrapersonal processes interact through cognitive and metacognitive processes, which include co-construction of meaning, co-regulation and co-reflection. With this architecture he is able to model the mutual influences.

The Havruta Learning Tradition

Havruta is a traditional Jewish method of studying texts, in which two partners (haverim) analyze and interpret a text together, often with the help of commentaries. Havruta is more

than just a way of reading; it is also a way of thinking and communicating, in which the partners challenge each other's assumptions, ask questions, and offer alternative interpretations (Kent and Cook 2014). Havruta encourages critical thinking, creativity, and dialogue skills, as well as a deeper understanding of the text and its implications.

This method, which has been practiced for hundreds of years, has been shown to be successful in developing students' sense-making skills for abstract learning, in schools (Holzer and Kent 2013) and in industry (Sapir, Drori, and Ellis 2016). The method entails the following steps. In each daily session, an instructor assigns a text to a dyad of students. The learning partners must interpret this text and challenge each other's interpretations in back-and-forth dialogue between them. The learning mainly occurs in the interaction between the two students, and the instructor is called upon only to clarify or resolve misunderstandings or disagreements between the students. Once the particular text is learned, another text is assigned to the same dyad, and a new cycle between the two students begins.

(Kent 2010), characterizes the Havruta setup as a set of three learning mechanisms that the dyads engage in, with each mechanism comprising two complementary learning activities: (i) dialogue, i.e., negotiating interpretations of the text by listening or articulating, (ii) attention management, i.e., wondering about different areas of the text or focusing on specific parts of the text and its context, and (iii) challenging or supporting perspectives. The right balance between the two activities within each mechanism is key to effective learning. The first learning mechanism, dialogue by listening and articulating, is the "main engine" of the learning process (Kent 2010). Such dialogue should proceed as a cycle of formulating one's own interpretation of the text—i.e., "listening" to where the text points and "labeling" one's understanding of it (Holzer and Kent 2013) —and then listening to your partner's interpretation, and subsequently articulating one's own interpretation through respectful argumentation. In Havruta learning, knowledge should be regarded as tentative, ready to be confirmed or disconfirmed by way of listening and articulating. Through this process, each partner forms a conceptualization—effectively a mental model—of the knowledge at hand, not unlike what Weick et al. (2005, p. 412) describe as "presumptive understanding through progressive approximations".

The second mechanism of Havruta learning, wondering and focusing, involves gathering context—with one or both partners thinking about and shifting between different parts of the text and its surrounding texts (context) and subsequently choosing to focus on a particular part and delve deeper into it. The act of moving attention to different parts of the context facilitates reciprocal learning by introducing new information for interpreting the original text but also allocating sufficient attention for deeper processing on a particular issue. (Jörg 2004) refers to this type of activity as "co-regulation": a meta-cognitive activity that emphasizes the need to agree on how to manage attention.

The third mechanism entails direct engagement with the partner's alternative perspective by challenging it, adapting it, or adopting it. You begin the dialogue by listening to your partner's articulation of his/her perspective and continue by supporting or challenging the perspective through respectful argumentation. Thus, a dialogue unfolds that rotates between the partners, in which one partner attempts to convince the other to accept a perspective, and the other partner responds with questions and arguments in support or in contradiction. Challenging or supporting is a way of practicing perspective-taking (Boland Jr, Tenkasi, and Te'Eni 1994).

Drawing from dyadic learning and Havruta for RHML

Both dyadic learning and Havruta share some common features that can inform the design of human-AI collaboration systems. First, they both involve a reciprocal relationship between the learners, in which they are both active participants and contributors to the learning process. Second, they both rely on dialogue as a key mechanism for learning, in which the learners negotiate meaning, provide feedback, and co-construct knowledge. Third, they both require a certain level of cognitive and social skills from the learners, such as working memory, inhibition, cognitive flexibility, perspective-taking, and communication skills.

Illustrative Scenarios

Given the above background and our stated position, three examples show how the human-human collaborative method might be applied to human-machine interaction.

(1) a human-machine poetry generation system – human/-machine for alternative perspective taking to generate original messages/metaphors. A machine might begin by learning rhyme and rhythm from human inputs; a human may encounter new metaphors suggested by the machine enriching their poetic abilities and appreciation; the machine may learn from rejected metaphors and improve its suggestions.

(2) text message classification as suspect/non-suspect – with human attention management to selected areas of the corpus. While a machine can perform an initial classification of texts, errors in those classification may teach a human how to use more informative or precise tags. Unexpected correct classifications may teach the human to consider tagging text patterns that were not previously apparent. (Te'eni et al. 2023,forthcoming)

(3) in-context sentence annotation – human use of high-level situational context versus machine linguistic context. When annotating sentences, humans and machines manage attention in different ways (Sen, Hong, and Xiaomei 2022). Humans rely on prior knowledge and intuition, utilizing their understanding of the world and contextual cues to guide their attention. On the other hand, machines allocate attention based on the context of terms in a sentence and their classification/prediction goals. Through RHML, machines can teach humans to be more objective-oriented, while humans can provide machines with a broader context that goes beyond the analyzed corpus, enhancing their understanding and decision-making abilities.

Conclusion

We believe that a promising direction for human-AI collaboration is to create systems in which machines and humans can act as dyadic partners or Havrutot (pl), by engaging in meaningful interactions and supporting their learning goals. This takes us beyond human-in-the-loop with its focus on control and safety, to human-in-the-learning-loop with a focus on human development alongside the machine. Such systems could leverage natural language processing and generation techniques to communicate between the learners, as well as machine learning and reasoning methods to provide relevant information and feedback. Moreover, such systems could adapt to the learners' needs and preferences, by using user modeling and personalization techniques. For example, a system could adjust its level of difficulty, pace, style, or tone according to the human learner's profile.

A number of key aspects cannot be inferred or adopted from the dyadic forms of human learning. Perhaps the most important and challenging is that of joint representations. Whereas humans can adopt, share, and develop shared representations through dialogue, common text, graphical depictions, and other forms of communication, machine readable representations of knowledge are (still) far more limited. This means significant effort is required to find and use the forms of shared knowledge representation accessible to both human and machine. Jörg's formulation of reciprocal learning is silent on the transfer of knowledge from one learner to another, concentrating on the dynamics of influence. Havruta presumes common text(s), language, location, and temporality of the study partners so has no challenges of knowledge representation that need to be addressed.

Other challenges that will need to be addressed in the move from human-human to human-machine include: How can RHML establish rapport and trust between learners? How can an RHML system handle ambiguity and uncertainty in natural language? How might a balance between providing guidance and allowing autonomy for the learners be established? Can we bypass the need for joint knowledge representations in human-machine learning dialogue?

Despite these challenges - or perhaps due to them - by drawing inspiration from the established human-human learning approaches of dyadic learning and Havruta, we may be able to design systems that can enrich and improve the human-AI collaborative experience. We dare say that in many situations in which humans and machines can learn together by doing and when human development is valued, RHML configurations should be considered.

References

Boland Jr, R. J.; Tenkasi, R. V.; and Te'Eni, D. 1994. Designing information technology to support distributed cognition. *Organization science*, 5(3): 456–475.

Holzer, E.; and Kent, O. 2013. *A Philosophy of Havruta: Understanding and Teaching the Art of Text Study in Pairs*. Academic Studies Press.

Jörg, T. 2004. A theory of reciprocal learning in dyads. *Cognitive systems*, 6(2/3): 159–170.

Jörg, T. 2009. Thinking in complexity about learning and education: A programmatic view. *Complicity: An international journal of complexity and education*, 6(1).

Kent, O. 2010. A Theory of Havruta Learning. *Journal of Jewish Education*, 76(3): 215–245.

Kent, O.; and Cook, A. 2014. Teachers as learners and practitioners: Shifting teaching practice through havruta pedagogy. *Religious Education*, 109(5): 507–525.

Sapir, A.; Drori, I.; and Ellis, S. 2016. The Practices of Knowledge Creation: Collaboration Between Peripheral and Core Occupational Communities. *European Management Review*, 13(1): 19–36.

Sen, W.; Hong, Z.; and Xiaomei, Z. 2022. Effects of human–machine interaction on employee's learning: A contingent perspective. *Frontiers in Psychology*, 13: 876933.

Te'eni, D.; Yahav, I.; Zagalsky, A.; Schwartz, D. G.; Silverman, G.; Cohen, D.; Mann, Y.; and Lewinsky, D. 2023,forthcoming. Reciprocal Human-Machine Learning: A Theory and an Instantiation for the Case of Message Classification. *Management Science*.

Vygotsky, L. S. 1978. Mind in society: The development of higher mental processes (E. Rice, Ed. & Trans.).

Zagalsky, A.; Te'eni, D.; Yahav, I.; Schwartz, D. G.; Silverman, G.; Cohen, D.; Mann, Y.; and Lewinsky, D. 2021. The Design of Reciprocal Learning Between Human and Artificial Intelligence. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).