https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–26
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Reciprocal Human-Machine Learning: A Theory and an Instantiation for the Case of Message Classification

Dov Te'eni,[a] Inbal Yahav,[a,*] Alexely Zagalsky,[a] David Schwartz,[b] Gahl Silverman,[a] Daniel Cohen,[c] Yossi Mann,[d] Dafna Lewinsky[d]

[a] Coller School of Management, Tel-Aviv University, Tel Aviv 6997801, Israel; [b] School of Business Administration, Bar Ilan University, Ramat Gan 5290002, Israel; [c] Department of Management, Bar Ilan University, Ramat Gan 5290002, Israel; [d] Department of Middle Eastern Studies, Bar Ilan University, Ramat Gan 5290002, Israel

*Corresponding author

Contact: teeni@tauex.tau.ac.il, https://orcid.org/0000-0002-5629-8614 (DT); inbalyahav@tauex.tau.ac.il, https://orcid.org/0000-0002-1513-017X (IY); alexeyza@gmail.com (AZ); david.schwartz@biu.ac.il, https://orcid.org/0000-0002-2125-2069 (DS); gsilverman@tauex.tau.ac.il, https://orcid.org/0000-0003-3937-6219 (GS); cohenda3@gmail.com (DC); yossimann1@gmail.com, https://orcid.org/0000-0002-9664-5385 (YM); dafnal1995@gmail.com (DL)

Received: November 14, 2022
Revised: May 14, 2023
Accepted: June 5, 2023
Published Online in Articles in Advance: November 14, 2023

https://doi.org/10.1287/mnsc.2022.03518

Abstract. There is growing agreement among researchers and developers that in certain machine-learning (ML) tasks, it may be advantageous to keep a "human in the loop" rather than rely on fully autonomous systems. Continual human involvement can mitigate machine bias and performance deterioration while enabling humans to continue learning from insights derived by ML. Yet a microlevel theory that effectively facilitates joint and continual learning in both humans and machines is still lacking. To address this need, we adopt a design science approach and build on theories of human reciprocal learning to develop an abstract configuration for reciprocal human-ML (RHML) in the context of text message classification. This configuration supports *learning cycles* between humans and machines who repeatedly exchange feedback regarding a classification task and adjust their knowledge representations accordingly. Our configuration is instantiated in Fusion, a novel technology artifact. Fusion is developed iteratively in two case studies of cybersecurity forums (drug trafficking and hacker attacks), in which domain experts and ML models jointly learn to classify textual messages. In the final stage, we conducted two experiments of the RHML configuration to gauge both human and machine learning processes over eight learning cycles. Generalizing our insights, we provide formal design principles for the development of systems to support RHML.

History: Accepted by D. J. Wu, special issue on the human-algorithm connection.

Funding: This work was supported by the Israel's Ministry of Defence [Grant R4441197567] and the Israel's Ministry of Science and Technology [Grant 207076].

Supplemental Material: The data files are available at https://doi.org/10.1287/mnsc.2022.03518.

Keywords: design science • human-machine interaction • reciprocal learning

1. Introduction

The combination of big data and artificial intelligence (AI) has dramatically changed the way people and organizations interpret data and make decisions (Grønsund and Aanestad 2020). It has become common, both in research and in practice, to develop machine-learning (ML) models for sense-making and decision-making tasks. These models are typically trained by human experts and subsequently operate autonomously, leveraging their capacity to efficiently process big data (that is, data that is too big to be handled manually) (Abbasi et al. 2018). Further developments in ML, coupled with even greater abundance of data—predicted to take place with the deployment of the Internet of Things and new forms of social media—are expected to heighten the extent to which tasks are delegated to machines (Baird and Maruping 2021). The current research is built on the well-established premise that fully autonomous ML performance, that is, "leaving the human out of the loop," produces suboptimal results in many settings. Take, for example, the task of classifying textual messages into a set of predefined labels, such as binary classification: Although several designs of ML classification models developed on the basis of theories and expert

knowledge have claimed to be effective in terms of classification accuracy (Abbasi and Chen 2008), such claims are generally limited to the models at their time of initial development, and evaluations carried out in later periods suggest poorer performance (Khashabi et al. 2020).

The idea of keeping the human in the loop (HITL), more specifically, keeping the human in the *learning* loop (So 2020a), encompasses cases in which machines continue to learn from humans. We wish, however, to take this view a step further to also encompass cases in which *humans* continue to learn from the machines they interact with. This learning, in turn, can enhance humans' capacity to provide valuable feedback for improving machine performance. Indeed, researchers increasingly acknowledge the benefits of such mutual human-ML (Seidel et al. 2018, Sturm et al. 2021, Van den Broek et al. 2021). Although the previous research recognizes the importance of HITL as a macro-level approach to designing ML systems, none provides specific details on how to design HITL at the microlevel to support learning by both machine and human, explicates the requirements of human-machine reciprocal learning processes, and outlines best practices for meeting them. This paper begins to address this need. Specifically, adopting a design science approach that integrates established theory on reciprocal learning between human partners (Jörg 2004), we develop an abstract configuration for a system that promotes joint continual learning of both human experts and ML models. To inform the development of our configuration and to demonstrate its performance, we developed a prototype called *Fusion* whose purpose is to classify textual messages. Fusion's initial design and preliminary evaluation were reported in (Zagalsky et al. 2021, blinded for review). In the current work we begin with the development of a theoretical basis, produce an extended design based on the theory and a design-science process, and conduct a series of experiments from two corpora in the cybersecurity domain to validate our approach.

1.1. Rationale for Keeping the HITL

More and more researchers and developers advocate keeping the HITL of AI-based decision making (Shrestha et al. 2019), in message classification (Duarte et al. 2018), and in other types of decision making such as medical diagnosis (Holzinger et al. 2021). Work addressing this topic encompasses multiple perspectives, integrating different rationales.

One stream of studies focuses on the benefits derived when *machines continue to learn from humans*. Some of the research in this vein is grounded in the efficiency-motivated perspective, which conceptualizes the human and the machine as cooperating partners rather than one being a tool for the other (Woods and Hollnagel 2006). This research suggests that human agents should continually audit and improve algorithms and data acquisition designs (Grønsund and Aanestad 2020). Ethics-motivated perspectives, in turn, suggest that continual human oversight of machines is necessary as a means of ensuring fairness in decision making (Enarsson et al. 2021). A third perspective stresses the importance of the HITL for continual training in cases where machines may fail to learn effectively (So 2020a). Limitations—both theoretical (Pearl 2019) and pragmatic (Khashabi et al. 2020)—on machines' ability to learn to cope with new situations require human experts to provide feedback in these situations and facilitate further learning (Zerilli et al. 2019). Combining the strengths of machine learning models with human intuition and expertise is crucial to the ongoing viability of an ML model, in particular when there is a need to identify contextual examples of out-of-distribution data (Groh 2022). The studies cited previously share a common understanding that the need to keep the HITL imposes design considerations that are not required in fully automated decision systems.

Our motivation in this work stems from yet another perspective, which focuses on the need for *humans*—particularly domain experts—*to continually learn from the machines with which they interact*. For instance, in classifying X-rays, the human-machine interaction offers opportunities for the expert to learn from the machine that would otherwise be difficult to provide (Abdel-Karim et al. 2020). In particular, the machine can highlight human classification errors, not merely to indicate erroneous applications of extant knowledge but also to indicate directions in which to generate new classification knowledge. Such knowledge can also be gained from collaborative decision-making processes. For example, in medical image analysis, after the ML models are trained, the expert can work jointly with the machine to interpret challenging medical images, a process that both improves the models and directs the expert's attention to information in the image that can improve the expert's subsequent classifications (McKinney et al. 2020, Budd et al. 2021).

Machine's learning from humans over time is exemplified in the evolution of human labeling practices for large datasets such as ImageNet, COCO, and Open Images. Outcome feedback generated by the machine is used to adjust the guidelines the labelers use, with the expectation that better human labeling will improve ML (Russakovsky et al. 2015). These examples highlight the significance of continual ML by keeping HITL to ensure learning by both machine and human. Indeed, our work on reciprocal learning extends and integrates these two learning processes.

Notably, the capacity to learn from machines is integral to experts' ability to teach machines in subsequent interactions. Experts lacking knowledge of new findings—knowledge that might be gained from machines—are

limited in their capacity to provide valuable feedback that might improve machines' subsequent performance. More crucially, in situations where experts are expected to oversee machines' decisions, a failure to learn from the machines' performance may lead to over-reliance on these machines and a failure to detect errors (Zerilli et al. 2019), which may lead to disastrous results (Omohundro 2014). Lebovitz et al. (2021) find that relying on AI outputs alone may severely limit learning and ultimately lead to a reduction of organizational expertise and suggest that when "both humans and AI tools operate in parallel, both retain the potential to learn and evolve, and there is the important potential to continue observing and scrutinizing the performance of both over time" (p. 1518), recognizing that it is a major challenge to achieve this form of interaction.

Our working assumption in this research is that humans should be kept in the *learning* loop—and the question is how the human-machine interaction should be designed. For both the machine and the human to learn from each other, a reciprocal model must be designed. We propose that such designs should be grounded in theory to ensure that the fundamental principles of reciprocal learning are realized in the human-machine interaction.

1.2. Developing a Configuration for Reciprocal Human-Machine Learning

In general, any configuration involving both humans and machines must ensure effective human functioning while maintaining the advantages offered by automation. Yet the specific requirements for a particular configuration depend on its underlying rationale and on the goals it aims to achieve (Suchman 2007, Amir et al. 2019). Herein, our goal is to create a general, abstract configuration that facilitates effective reciprocal learning of both humans and machines—where the emphasis is on optimizing the learning processes that *both* parties undergo, such that other aspects (e.g., the machine's autonomous performance in periods when neither party is undergoing active learning) are secondary. We call the proposed configuration a *reciprocal Human-ML (RHML) configuration*.

Our research methodology is design science, involving a multidisciplinary team of researchers. We consider the theory of reciprocal learning (Jörg 2004) to be kernel knowledge and the foundation of our approach. According to Jörg's theory, reciprocal learning is a process of cocreation that entails "a complex of self-generative, self-sustaining processes of mutual 'bootstrapping' with potentially nonlinear effects over time" (Jörg 2009, p. 1). In line with this idea, a key assumption of the proposed RHML configuration is that the human and the machine learn from each other to solve problems, analyze, explain, and judge. They learn in repeated cycles of two intertwined learning processes, creating a representation of the knowledge learned.

We further develop an instantiation, Fusion, that demonstrates the use of the RHML configuration in learning to classify textual messages. In this context, our RHML configuration operates according to the following principles. Human experts establish a ground truth as a basis for training the ML classification models (though, theoretically, training need not involve human agents). Once trained, the ML interacts with a domain expert in several cycles of reciprocal learning. In each cycle, the machine classifies multiple messages and generates feedback with an analysis of the classifications, for example, compatible and incompatible classifications. From the feedback, the expert then learns new considerations for classifying messages in the future and provides these new classification insights as feedback to the ML for further improvement. The machine updates its classification models and enters a new cycle to classify new messages. These cycles of reciprocal learning continue until a point of saturation. The machine is then left to operate autonomously based on what it has learned until a new set of learning cycles is initiated periodically or after some significant change in the environment.

This form of human and machine learning in cycles leverages the two parties' complementary capabilities. In particular, the machine applies and adapts classification models to big data efficiently, and the human identifies and applies messages' context to determine their meaning (a capability we discuss in more detail later). In this setup, unlike in some of the human-machine configurations described in the previous section, not only does the human reinforce ML with feedback and new insights, but the machine also reinforces human learning with appropriate machine-generated feedback.

Our proposed configuration addresses several core challenges associated with the integration of humans and machines and harnessing their complementarities for the purpose of reciprocal learning. The first relates to the need to allocate learning activities between the human and the machine. In contrast to reciprocal interactions between two human partners, in which activities can effectively be carried out symmetrically by both parties (as they have similar cognitive abilities), RHML involves distinct activities for the human and the machine, corresponding to their different capabilities. These activities must be identified and allocated appropriately—a process that can become arbitrarily complex (Fügener et al. 2021). In this case, the long-standing relative performance criterion for allocating activities between humans and machines for the purposes of achieving high task performance (Fitts 1951) must be expanded to more complex and fuzzy criteria that also consider the transfer of control and responsibility from machine to human (Schwartz 1995). The

tasks we consider are complex. The immense challenges that task complexity places on automation potential and opacity are raised in detail by Vimalkumar et al. (2021), who provide a framework to differentiate fairness concerns across different task types. We believe that RHML can contribute to resolving challenges of opacity and increase automation potential for complex tasks. The second challenge relates to communication. The need for mutual understanding between learning partners—a challenge even in communication between humans—matters even more in human-machine communication (Suchman 1987). The complexity of human-machine communication is particularly high when facing intelligence-intensive tasks, such as gaining insights, mindful judgment, creativity, and contextualization compared with routine, structured, or programmed tasks.

To address these challenges, we carry out (i) a functional analysis that identifies the specific learning activities that each party must be able to perform and (ii) a communication analysis that establishes requirements for effective communication between the two parties (see Section 3). Together, these analyses reveal a crucial requirement for the functionality of an RHML configuration: inclusion of explicit representations of knowledge that are simultaneously compatible with both machine and human processing, facilitating the interactions between parties as well as their individual learning activities.

Our research methodology for studying RHML builds on a mix of previous recommendations for design science research (DSR; Hevner et al. 2004, Sein et al. 2011, Gregor and Hevner 2013, Lee et al. 2015) and it produces three artifacts that encapsulate the design knowledge gained: a theory artifact, a technology artifact, and design principles. We worked in parallel on defining the abstract RHML configuration and on implementing Fusion. To achieve our goals, we formed an interdisciplinary multi-university team of nine researchers with different areas of responsibility and expertise, as elaborated in the following subsection.

We make three primary knowledge contributions, corresponding to the three design science artifacts. First, as a theory artifact, our RHML configuration is an exaptation of established theory on reciprocal learning; specifically, it translates the concept of reciprocal learning in human dyads—presented in the original work of Jörg (2004, 2009)—to the context of human-machine dyads. Second, Fusion (our instantiation of RHML) is a novel technology artifact that contributes an improvement in the way human experts and machines can interact to solve text classification challenges over time. Third, we propose a formal design principles artifact for supporting RHML configurations, providing new guidance for settings in which machines and humans need to learn from each other over time.

2. Overview of Methodological Approach: Design Science

Our research was conceived to follow the "Systematic High-Impact Research" model of Nunamaker et al. (2017) and was conducted under the auspices of a national research center to address a real-world problem. Our approach is anchored in DSR, which is aimed at studying the development of new artifacts, be they theoretical, technological, or methodological, as a means for gaining knowledge. We follow a "proof of concept" approach (Iivari 2015) in which a system is constructed to address a general problem and then instantiated as a test of the design theory. In our context, this approach translates into development of a theory artifact (the abstract RHML configuration) and its instantiation as a corresponding technology artifact (Fusion), used to test RHML in two case studies of text classification. Although this paper presents them separately, the theory and technology artifacts were developed as intertwined and iterative efforts, as summarized in Table 1. In this way, we achieve a synergy between the human and machine goals that coexist in RHML, corresponding to the Gestalt mode of DSR as described in Adam et al. (2021).

***Stage 1: Formulate the research problem*:** We identified basic theoretical tenets of reciprocal learning from prior work on human dyads (Jörg 2004, 2009) and translated these into functional requirements for our RHML configuration while also identifying additional

Table 1. Our Four-Stage DSR Process, Culminating in (i) a Theory Artifact (RHML Configuration); (ii) a Technology Artifact (Fusion), and (iii) a Design-Principles Artifact (Guidelines for Developing RHML)

Stage	Abstract RHML configuration (theory artifact)	Instantiated Fusion (technology artifact)	
1 Formulate research problem	Formulate RHML	Specify design to support RHML configuration	Section 3
2 (a) Build		Develop the software	Section 4
(b) Use		Users (domain experts) use Fusion to classify messages	
(c) Evaluate and reiterate		Test functionality and usability and adjust Fusion	
3 Evaluate reciprocal learning in experiments	Measure quality of ML and Human learning	Adjust Fusion	Section 5
4 Generalize	Augment RHML and set its boundaries	Formulate design principles	Section 6

challenges that must be addressed in implementing human-machine reciprocal learning.

Stage 2: Build-Use-Evaluate an instantiation of the RHML configuration (Fusion): We worked on two cases with their respective corpora taken from two different Darknet forums, each corpus with its corresponding domain experts. Stage 2 was repeated as four DSR iterations.[1] We carried out two iterations on the first corpus and then introduced the second corpus for the subsequent two iterations. We worked on a cascade of corpora, learning, and interchanging between cases as we proceeded. With the results of each evaluation in stage 2, and based on the lessons learned, we *reiterated* and *redesigned* both Fusion and our abstract RHML configuration. For example, we adjusted the workflow of the domain experts and the selection of qualitative techniques they were instructed to use; moreover, additional types of feedback from the machine to the human were designed. Additional modifications included introducing more functionality in Fusion and improving usability (Zagalsky et al. 2021).

Stage 3: Evaluate reciprocal learning in experiments: After four iterations of Build-Use-Evaluate (stage 2) followed by corresponding redesigns, we conducted a summative evaluation of reciprocal learning in the form of two repeated-measures experiments in which we compared a baseline ML classification model with the RHML configuration. Each experiment paired a human domain expert with corresponding machine models for a series of learning cycles facilitated by Fusion. At the end of each learning cycle, the experts completed a questionnaire about their perceived learning. This allowed us to study RHML in action and evaluate the human and ML.

***Stage 4: Generalize the RHML configuration and formulate design principle*:** After concluding the DSR iterations (of stage 2) and learning evaluations (stage 3), we generalized the lessons learnt into a final abstract RHML configuration. As part of this stage, we derived explicit design principles for the creation of RHML configurations. The theoretical model and the design principles are distinct DSR artifacts and make distinct knowledge contributions.

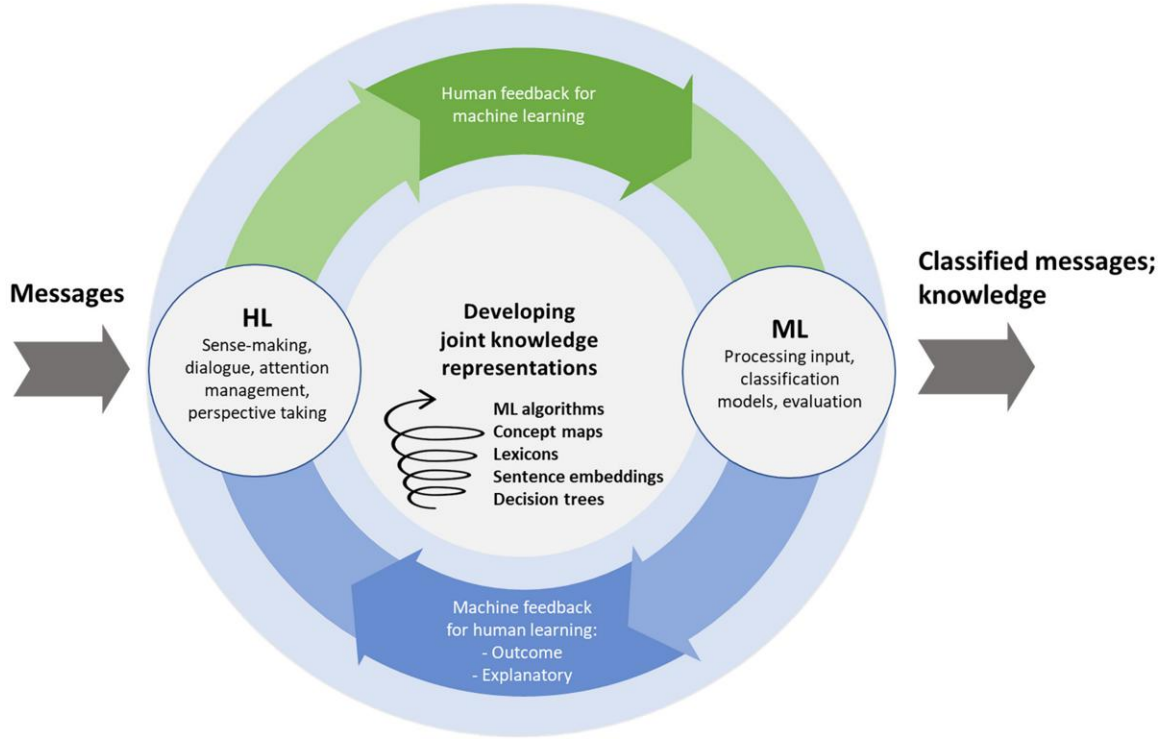
To execute the stages outlined previously, we formed a team of nine researchers composed of (I) a data science subteam responsible for developing ML algorithms; (II) domain experts who developed explicit representations of their knowledge and participated in user studies; (III) a qualitative research-methods expert; and (IV) a design subteam that formalized the human-machine processes and oversaw the development of Fusion. In addition to the academic team, a collaboration was established with a cybersecurity software company that specializes in analyzing communication in the Darknet, which served to anchor the real-world demands of our chosen problem domain.

3. Theory Artifact: Develop an Abstract RHML Configuration

In this section, we describe the development of our theory artifact: the abstract RHML configuration. We begin with the basic premise that the essence of reciprocal learning between human and machine is the interaction between their respective learning processes, namely the human learning (HL) and the ML. These two learning processes are distinct from each other, as the two parties possess different capabilities. For example, as we will elaborate, ML text-classification models and human sense-making processes differ in their capacities to use context for making sense of texts, in the way they manage attention to information, and in their ability to consider and adopt alternative perspectives. Yet in RHML, the two learning processes are intertwined, with one feeding into the other in cycles, which progress toward joint development of classification knowledge. Any configuration that enables RHML must support the HL and ML learning processes separately while simultaneously supporting the interaction between them in a manner faithful to the principles of reciprocal learning. To derive the principles of reciprocal learning in human–machine dyads—and given that no formal paradigm for such learning exists, to our knowledge—we draw from a theory of reciprocal learning in human dyads (Jörg 2004, 2009), coupled with concrete characterizations of the mechanisms involved in such learning between students (Holzer and Kent 2013). We then build on these principles to formulate the specific problem of human–machine reciprocal learning. Finally, we present our abstract RHML configuration, which addresses this problem. As noted previously, in practice, this configuration is informed not only by theory but also by insights gained from parallel development of Fusion (as outlined in Table 1).

Figure 1 describes the essence of the RHML configuration. HL and ML represent the different learning processes performed, separately, by humans and machines. Each learning process has its distinctive functions such as sense-making (human) and classification (machine). Two distinct tailored feedback loops indicate the interaction between human learning and machine learning. One feedback loop is designed to support the learning of humans with feedback generated by the machine, and the other is feedback based on human insights designed to enhance the learning of machines. In the center of the diagram, we see an evolving set of joint knowledge representations indicated by an expanding spiral that grows continually to include more interactions, more information, more perspectives, and more confidence in putting knowledge to action (Nonaka et al. 2000). Joint knowledge representations act as both a product of reciprocal learning and an enabler of reciprocal learning.

Figure 1. (Color online) Learning Cycle in RHML



Although the role of product is intuitive and expected, it is in the role of enabler where the RHML joint knowledge representations make their critical contribution. With richer and more varied forms of joint knowledge representation that can be shared between human and machine comes greater reciprocal learning opportunities. Bracketing the reciprocal learning cycle are inputs and outputs which, respectively, in this study, are messages and classified messages.

3.1. Theoretical Background (Kernel Knowledge): Human–Human Reciprocal Learning

Our concept of RHML builds on the theory of learning in human dyads of Jörg (2004, p. 1; 2009), which 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 coproducing 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 meta-cognitive processes, which include coconstruction of meaning, coregulation, and coreflection. With this architecture, he is able to model the mutual influences. Jörg's formulation of reciprocal learning is silent on the transfer of knowledge from one learner to another, concentrating on the dynamics of influencing one another.

A revealing demonstration of the specific mechanisms involved in reciprocal learning can be found in a process called *Havruta*, a traditional Jewish method of learning in dyads. 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 et al. 2016). The method entails the following steps. In each daily session, an instructor assigns a text to a dyad of students. The two 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 on 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.

As discussed by Kent (2010), the Havruta setup can be characterized as a set of three learning mechanisms that the dyads engage in, with each mechanism

comprising two complementary learning activities. The three mechanisms are (i) dialogue, that is, negotiating interpretations of the text by *listening or articulating*, (ii) attention management, that is, *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, hence, the need to coordinate and control the mechanisms and activities. 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—that is, "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 fact, one can imagine two dialogues: one dialogue between the student and the text, and the other dialogue between the students.

The dialogue between partners enabled by listening-and-articulating is the platform for the learning cycles because it provides the opportunity for incremental learning by testing old beliefs and interpretations in light of new data, new perspectives, and new contexts that lead to new or refined conceptualizations in new cycles. The combination of listening and articulating ensures that sufficient attention is devoted to generating or refining the conceptualizations. At any moment, 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; Weick et al. (2005, p. 412) describe such conceptualizations 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 et al. 1994).

Among the three mechanisms discussed previously, the latter two—namely, wondering/focusing and challenging/supporting—are particularly effective in facilitating the abstract learning process, as they encourage the learner to (i) think beyond the immediate context of a text, (ii) go against one's own perspective, and (iii) abstract from concrete observations to a more general conceptualization.

3.2. Problem Formulation: From Human-Human Reciprocal Learning to Human-Machine Reciprocal Learning

Borrowing from the human learning theory of Jörg theory and the Havruta practice to the human-machine context is challenging, as discussed briefly in Section 1. We envisage, like Jörg, a similar architecture of two intra-agent learning processes that interact (Figure 1), but unlike Jörg we must assume that the two intra-agent and respective meaning spaces are different, one is natural and the other an artifact, and furthermore, reciprocity is designed rather than reliant on social principles. In the RHML configuration, therefore, the learning and knowledge transfer mechanisms are modeled differently. Additionally, the different human and machine learning models open the possibility of allocating activities according to relative advantages, which was not part of Jörg's theory. We identify three challenges in the transition.

The first challenge resulting from the distinct capacities for information processing and sense-making, is the need to identify the different activities involved in the RHML learning processes and to allocate them appropriately between parties. The second challenge relates to the fact that the complexity of human-machine communication is greater than that of human-human communication, partly because humans and machines represent and apply knowledge differently. To address these challenges, we carried out a functional analysis and a communication analysis. We subsequently discuss a third challenge: the need to develop joint knowledge representations compatible with both human and machine.

3.2.1. Challenge 1: Identification and Allocation of Learning Activities (Functional Analysis). A functional analysis identifies and allocates the activities that different components of a system should execute; in our case, these activities included the specific learning activities that the human and machine must engage in, as part of RHML. Our functional analysis did not address activities occurring outside the learning per se, such as the machine's autonomous performance after the learning cycles. Our analysis was grounded in the

premise that—like human-human reciprocal learning that involves a person's private dialogue with the text, as well as the dialogue between partners—in RHML, the human and the machine engage separately in some learning activities (active in HL *or* ML), as well as the human-machine interaction needed to coordinate reciprocal learning by perspective taking and attention management (enabled through the feedback loops between HL and ML).

Our functional analysis differs from traditional allocations of activities between machine and human in which the goal is to divide labor in a manner that capitalizes on the agents' relative advantages in achieving performance on a focal task (Fitts 1951, Ip et al. 1990). For text classification tasks, for example, common allocations aimed at achieving classification accuracy include the following: (1) for data labeling, humans assign a category to texts in a training set and then provide feedback to the machine by relabeling false ML-based classifications, and (2) for feature engineering, humans generate or select the features with high prediction power, commonly in a trial-and-error fashion, whereas machines apply the classification models to high-volume data (Marcellino et al. 2020). In our case, however, the goal of the allocation is not only to achieve ML classification accuracy but also to develop the human expert's knowledge. This dual goal may affect the optimal allocation of activities between parties by implying tradeoffs when selecting an optimal ML model (e.g., sacrificing some level of classification accuracy in favor of improved human comprehension and learning; Raisch and Krakowski 2020).

A key consideration in our functional analysis relates to humans' and machines' different capabilities with regard to the use of context in the learning process. For humans, attention to context and the use of context for interpretation are central to information processing in general and to effective learning specifically (Goldstein 2014). HL rotates, often as trial-and-error, between a highly contextualized interpretation of a message to a more abstract, and necessarily decontextualized, view of the message (Roth 2005, Son and Goldstone 2009, Zamani et al. 2021). People typically rely on two types of context to interpret a focal text. The first is linguistic context, broadly defined as written or spoken information that appears in conjunction with the text. The second is situational context (physical context), which refers to the general knowledge that a person has of the world that is relevant to the subject (Fromkin et al. 2011), for example, knowing that a message was written outside normal activity hours. People use linguistic context to identify possible meanings for a text and to clarify the correct meaning in cases of ambiguity; they also share context to clarify the meaning to others as needed (Schwartz and Te'eni 2000). Situational context also plays a key role in individuals' capacity to make sense of messages (Diedrich et al. 2011). Knowledge about the larger corpus from which a text was extracted and information gained from outside the corpus can inform the analyst's overall perspective and, correspondingly, his or her interpretations of the text. For example, analysts who come from different professional backgrounds or have different native languages may hold different perspectives that lead to different interpretations and conflicting classifications of the same text (Diedrich et al. 2011). Thus, humans' capacity to leverage context can enrich their sense-making and improve learning (Weick et al. 2005).

Unlike people, machines commonly rely only on linguistic context in the sense that terms are processed as an integral part of the text that surrounds them (Khashabi et al. 2020). Specifically, context is commonly leveraged from large textual corpora via embedding techniques (Mikolov et al. 2013), in which context of terms is derived from their surrounding terms. Recent models further incorporate attention mechanisms to automatically model the relationship between proximal term embeddings, thus emphasizing the local context of terms (Devlin et al. 2018). These embeddings may miss the broader context of a text, which manifests a hidden "internal conceptualization in the human mind" called a "meaning space" (Khashabi et al. 2020, p. 3). In contrast to this meaning space, which is free of noise and uncertainty, the linguistic space of the textual corpus (on which embedding techniques operate) is noisy and incomplete and therefore deficient in interpreting the message correctly. Recognizing this deficiency, recent attempts have begun examining ways to automatically incorporate layers of context into language models (Gardner et al. 2018, Wang et al. 2019). Alongside these attempts, there is growing agreement in the general ML research community that not all text classification tasks can be delegated to ML (Shrestha et al. 2019, Grønsund and Aanestad 2020). In our research, we rely on the human expert to provide situational context.

Our functional analysis identified the following learning activities to be allocated to human or machine (Figure 1):

(1) *Sense-making activities* of the ML and the HL processes; these include (a) examining a text, articulating its meaning, and categorizing it, (b) identifying relevant context, focusing on it and using it, and (c) taking perspectives;

(2) *Coordination and control activities* enabled by the two feedback processes between HL and ML, which include (a) listening and articulating to the other party and one's own understanding, (b) moving and focusing the dyad's attention, and (c) sharing perspectives.

The learning process may take time, a fact that is not conveyed in Figure 1. Although machines are usually

assumed to work quickly, human sense-making may occur over a protracted time. Processing the feedback from the machine may take weeks.

3.2.2. Challenge 2: Identifying Mutual Communication Requirements (Communication Analysis). The RHML configuration requires coordinated learning and mutual intelligibility (Suchman 2007). Even in human-human reciprocal learning, communication is complex—where communication refers both to the dialogue with the text and the dialogue between partners (Weick et al. 2005). In the human-machine case, communication is further complicated by the differences between their respective models. Figure 1 shows the communication as the two feedback processes flowing between the ML and HL processes. Each feedback process has its role and form.

Specifically, the feedback communicated by the machine to the human is intended to enable sense-making (Baralou and Tsoukas 2015). As noted in the discussion of our functional analysis, human sense-making involves moving back and forth between abstract and contextualized information; this process implies that the content of the feedback describing the context should vary in its level of detail. When the communication gap between the machine and the human is greater (i.e., less common ground), more elaborate feedback is needed to ensure understanding (Te'eni 2001).

In line with the requirement for the human to receive communications that vary in their use of context, we identify two forms of feedback that the machine should provide the human expert. The first is *outcome feedback*, which indicates whether a classification was correct or not, or, in the case of multiple predictions, the overall accuracy of the set. The second form of feedback is *explanatory feedback*, which includes context that explains why a classification is correct or incorrect, and in the case of multiple predictions, it may explain low accuracy due to some bias in estimation. Explanatory feedback, compared with outcome feedback, includes more concrete and detailed context around a core message (e.g., specific examples of true or false classifications). Research on learning with multimedia has shown that combinations of outcome (corrective) feedback and explanatory feedback ensure deeper learning compared with either type of feedback on its own (Moreno 2004). More generally, computer support can combine multiple layers of context and deliver it interactively so that only relevant feedback is given at the right time in a manner that is easy to understand and use effectively (Katz and Te'eni 2007, Lipton 2018, Rudin 2019).

3.2.3. Challenge 3: Achieving Joint Knowledge Representations. To fulfill the functional and communication requirements for RHML, the human and the machine must both be able to produce *external representations* (external to the human brain) of their classification knowledge that, on the one hand, can be used by the machine to classify text and, at the same time, support the human learning process. In practice, such representations may take on different forms, defined with different formalisms and residing in separate physical data files or as part of the code. The representations may include multiple types of structures such as concept maps, knowledge graphs, decision trees, decision tables, lexicons, algorithms and more.

The spiral in Figure 1 denotes the evolving nature of the knowledge generated by reciprocal learning. The spiral of knowledge-creation pauses to take action and resumes to upgrade itself continually with increasingly richer contexts and more interactions (Nonaka and Takeuchi 1995, Nonaka et al. 2000). In the HL process, new information is assimilated and accommodated in concept-relationship schemas in human memory (Novak 2010). In our research we aim to capture changes in the knowledge gained by reciprocal learning in joint knowledge representations.

3.3. RHML Configuration

Based on the theory outlined previously, combined with our functional and communication analyses, we propose the abstract RHML configuration (Figure 1). It constitutes a continual reciprocal learning scenario for the context of text classification: (1) expert generates a conceptualization; (2) machine accepts feedback from the expert (elements of the conceptualization); (3) machine runs classification models; (4) machine returns feedback to the expert; and (5) expert sees the feedback and manually revises the conceptualization.

Here we characterize the three components of the configuration: HL and ML processes, feedback for coordination and control of learning activities, and joint knowledge representations. In practice, the ongoing cycles are preceded by initialization operations that include building and organizing the corpus of messages, establishing a ground truth for supervised models, and designing ML algorithms. (This configuration should be seen as tentative in the sense that new technologies will afford new functionalities and new activity allocations.)

HL and ML Processes: The HL and ML processes of the RHML configuration have as inputs the data (the corpus of textual messages) and joint knowledge representations. The processes, composed of individual learning activities, interact with each other through human-generated and machine-generated feedback loops. The agents performing the processes are a domain expert engaged in sense-making and a machine operating with a set of ML classification models, designed and redesigned by a data scientist.

Coordination and Control of Learning Activities Through Feedback: The RHML coordination and control mechanisms are captured in the feedback loops. These mechanisms include managing attention by determining the context on which to focus, repeating reciprocal learning cycles to enhance sense-making, shifting between outcome feedback and explanatory feedback, and generating and analyzing alternative perspectives.

Joint Representations of Classification Knowledge: The HL process generates a conceptualization of the data, which is represented as concept maps with decision trees, along with an accompanying set of decision rules. Additionally, linguistic dictionaries (lexicons) map the concept maps and decision rules to texts. The ML process, overseen by the data scientist, produces a vector representation on the textual input informed by the rules and dictionaries, and an input to the algorithm. Output representation includes a usefulness score on each value of the original rules and dictionaries, suggestions on how to extend the dictionaries, and detailed (per-message) performance evaluation.

4. Technology Artifact: Fusion

This section describes the development of our technology artifact, Fusion. As discussed in Section 2, development included four iterations of building and adjusting, using, and evaluating Fusion (stage 2 in Table 1). We first introduce our case studies that serve as the basis for the Build-Use-Evaluate iterations. In the section titled "Build Fusion," we describe key elements of Fusion's final structure and interface and demonstrate how Fusion realizes the core system functionalities discussed in Section 3. Next, in the section titled "Use and Evaluate Fusion," we elaborate on the procedures followed to evaluate Fusion's functionality and usability with our two case studies

4.1. Case Studies

We constructed two case studies in which domain experts and ML models jointly classified texts taken from corpora in the cybersecurity domain. The corpora were chosen at the beginning of the project and were motivated by our industry partner specializing in Darknet communication analysis and threat detection. We used our industry partner's API to collect the data. The data we used involved several inherent challenges—including strong data imbalance toward nonsuspect messages, hard-to-distinguish language use in both suspect and nonsuspect groups, the use of slang, and the unstructured nature of the data.

4.1.1. Case Study 1—Hidden Answers: Identifying Illicit Drug Transaction Messages. *Hidden Answers* is a Darknet version of Q&A sites such as Stack Exchange and Reddit, where users can post questions about effectively any topic (without censorship). This is a diverse forum with very broad themes, and it operates in English, Spanish, Portuguese, and Russian. Allegedly, users can also ask crime-related questions, such as "Where can I buy drugs?," "Is the site illegal?," and "Where to buy guns and fake IDs on the Deep Web?" The interface is similar to the interfaces on Reddit, in that it is centered around questions organized by tags and offers search functionality.

DE1, the domain expert assigned to Hidden Answers, is a specialist in social discourse analysis and cyber ethnography. His goal was, through qualitative sentiment analysis, to classify as "suspect" those messages that talk about use, solicitation, selling, and purchasing of illegal drugs. We collected a total of 5,337 messages (containing questions, answers, and comments) for this case study from March 8, 2018, until April 25, 2018. At this point, DE1 built an initial conceptualization based on 513 messages: inspecting the messages, assigning categories to relevant messages through qualitative content analysis and then building the lexicon through secondary structural analysis. To avoid overfitting, only when this process was done, the rest of the data (4,824 messages) were classified to serve as ground truth when using Fusion. After removing messages that could not be classified, we ended up with 5,285 messages.

4.1.2. Case Study 2—BitsHacking: Identifying Professional Hackers. *BitsHacking* is an English-language cybercriminal Darknet forum operating since 2012. It is known as one of the most popular carding sites; however, it also includes hacking and security, cracking, dump sharing, tutorials on various topics, and hacking competitions.

The classification goal in this case was to differentiate between professional hackers and amateur hackers. Specifically, the goal was to classify as "suspect" messages indicating a professional hacker, that is, a person with the knowledge and means for conducting a harmful attack. We collected a total of 3,242 messages (containing questions, answers, and comments) for this case study from December 28, 2019, until January 26, 2020. As in the first case study, a domain expert (DE2) built an initial conceptualization based on 543 messages. However, because of data imbalance reflected by a low number of suspect messages, the domain expert increased the initial data set to 1,575 messages (containing 59 suspect messages). To avoid overfitting, the rest of the data (1,668 messages) were classified by an external domain expert (DE3) specializing in cybercrime to serve as ground truth during the iterative process within Fusion (Table 2). During this process, both DE2 and DE3 conferred about the classification process and discussed which examples they considered to be suspect or not (as part of an interrater reliability protocol established

Table 2. Descriptive Statistics

Case study	Data size	Class distribution	Unique users	Word count
Hidden answers	5,285 messages (513 initially)	196 suspect (3.85%)	1,166	57.43 average
BitsHacking	3,242 messages (1,575 initially)	78 suspect (2.46%)	344	198.59 average

Notes. Data size is the number of messages in the corpus (parentheses indicate the number of messages used in initial training). Class distribution indicates how many messages were 'true suspects' (and their percent out of all messages).

ahead of time). DE2 is a cybersecurity and web intelligence expert, specializing in identifying cyberterrorism and the use of Darknet and social media by nonstate actors. He consults for industry business-to-business (B2B) cyberintelligence firms and previously served as an expert consultant on online radicalization for the Organization for Security and Cooperation in Europe (OSCE). DE3 is a senior threat intelligence researcher, whose field of expertise is cyberintelligence including collection, research, and analysis of criminal activities in the cyberspace and Deep Web environments.

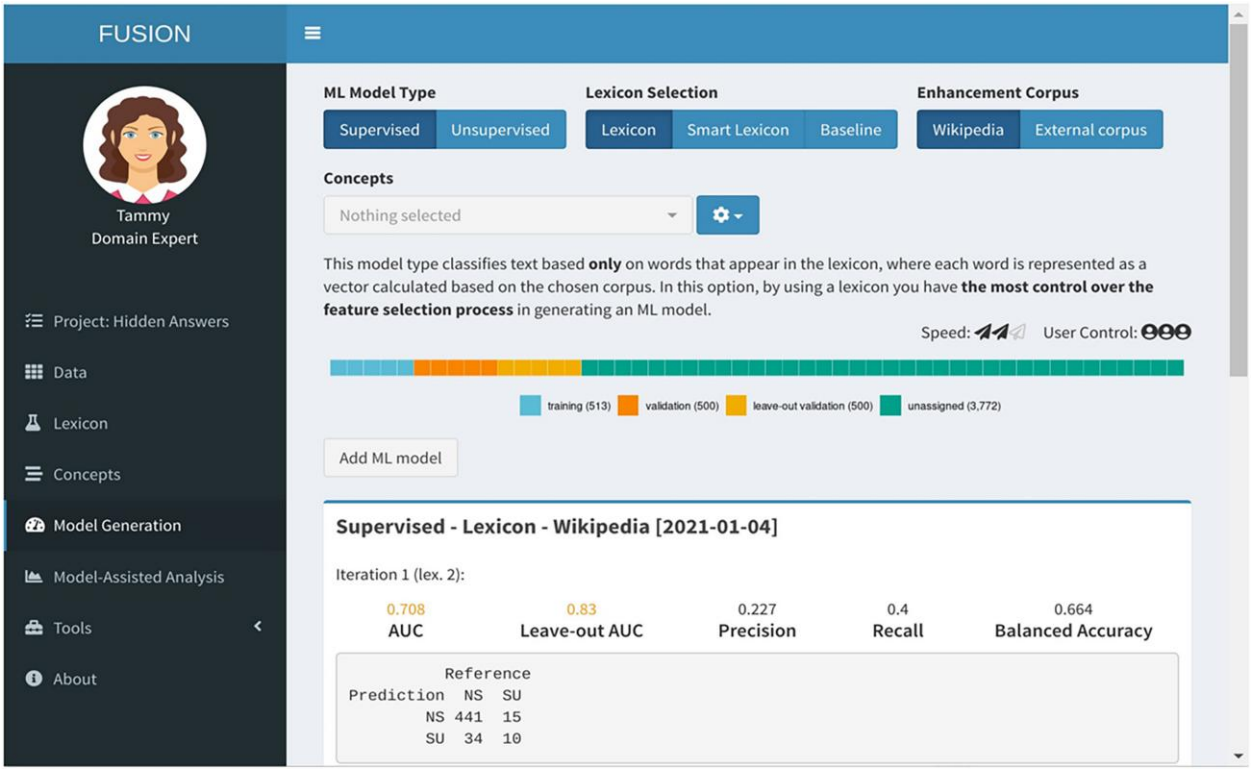
4.2. Build Fusion: Overview of the Interface

Fusion, a dynamic webapp,[2] introduces the idea of reciprocal learning into extant classification systems, and it offers many of the core capabilities that are common in state-of-the-art systems. These capabilities include the following: (1) GUI-based model construction—a user-centric approach in which the user outlines the data analysis process (examples of tools that incorporate this capability include deepcognition.ai, dataiku.com, orangedatamining.com, rapidminer.com); (2) data visualization capabilities (e.g., bigml.com); and (3) automatic data-driven ML—a machine-centric approach with minimal user intervention, aimed at lower-tech users. Fusion is written in R (R Core Team 2021) and uses the shiny framework for R (Chang et al. 2021).

Figure 2 presents a screenshot of Fusion's user interface. The menu on the left-hand side of the screen contains items corresponding to the key functionalities supporting the processes described in our RHML configuration (Figure 1). The menu item Data initializes the project's corpus. The menu items Lexicon and Concepts allow the user to influence the ML process, and Model Generation controls the selection and design of ML models. Last, Model-Assisted Analysis supports sense-making based on machine-to-human feedback.

The Model Generation item allows the user to select a combination of model type (supervised or unsupervised),

Figure 2. (Color online) Model Selection and Generation Screen with Results of a Selected ML Model



lexicon, and enhancement corpus. Specifically, this item provides the following options, shown on the top area of the screenshot in Figure 2:

- *ML Model Type: Supervised or Unsupervised:* The supervised models employ word2vec to represent the textual input and lasso regression via glmnet (Friedman et al. 2010) to generate the model. Unsupervised models refer to sentiment scoring methods (Pang and Lee 2004), implemented as follows: classify a message by examining how many terms (words, phrases) from the message are labeled "suspect" versus "nonsuspect," normalized by the size of the lexicon.
- *Lexicon Selection: Lexicon, Smart-Lexicon or Baseline.* Lexicon and Smart-lexicon options use a lexicon file for feature selection, essentially "focusing the machine's attention" on important words and filtering out noise. Lexicon models use the lexicon as given, while Smart-lexicon models first enhance the lexicon with semantically similar words (based on their embeddings) and then generate the model. Selection of the Baseline option implements a model that does not rely on reciprocal learning, meaning that the user has no control over feature selection. The baseline model classifies texts by considering all meaningful words, where each word is represented as an embedding vector calculated with the enhancement corpus, and each sentence is computed via an average embedding of its words. We use this model as a benchmark for noninformed machine learning.
- *Enhancement Corpus Wikipedia or Other*: Fusion uses a built-in enhancement corpus based on a portion (100 MB) of the English version of Wikipedia. For specialized cases where specific terminology is used—for example, online forums in which slang is prevalent—the user can provide a custom enhancement corpus.

Together, these characteristic selections indicate the type of model to be generated. In total, there are currently five possible model types, each with its own characteristics and unique approach for classifying text: (1) *unsupervised-lexicon*, (2) *unsupervised-smart-lexicon*, (3) *supervised-lexicon*, (4) *supervised-smart-lexicon*, and (5) *supervised-baseline*.

The Concepts item on the menu enables the domain expert to introduce classification rules derived from a concept map, one of our primary representations of human knowledge; these rules are transformed into a machine-readable feature, which can be added to the embeddings features. An example of a rule can be "low probability of suspect if lack of anonymity," where lack of anonymity will be formulated as disclosure of identifying information such as phone number and email address.

In what follows, we discuss the various elements of Fusion's interface in depth, focusing on the manner in which they support the core functionalities of our RHML configuration, and mapping the functionality and its design rationale to emerging design principles, which are summarized formally in the Discussion section.

4.2.1. Supporting Joint Knowledge Representations (Emerging Design Principle 1).

Reciprocal learning necessitates the use of joint knowledge representations to classify and generate new knowledge, but at the same time, the learning is shaped and constrained by these representations, for example, constrained to certain perspectives. As the processes of creating and using knowledge and the structure of knowledge are intertwined, both are described in this section, beginning with structure and continuing with the processes.

In the abstract formulation of RHML, the joint representations were seen as one repository, generated and shared by machine and human. Implementing the joint representation in Fusion surfaced the challenges of designing appropriate data structures accessible to human and machine processing. Currently, Fusion represents knowledge in two forms, namely, lexicon and concepts (respectively, the third and fourth menu items in Figure 2).

Table 3 shows an example of a lexicon, which was developed through qualitative content analysis and structural analysis from Case Study 1. A lexicon term can be a word, a phrase comprising multiple consecutive words, or word patterns, such as "X AND Y." The lexicon includes the domain expert's hierarchy of categories of terms (representing concepts) and their mapping to word strings in the corpus. The major categories are mapped to the target classes, which in our case, are suspect (SU) or nonsuspect (NS). The system uses this mapping to interface between the domain expert's conceptualization and a machine-readable (and testable) input. Fusion allows the ML models to directly access the lexicon, and for the

Table 3. Example of a Lexicon of Suspects (SU) and Nonsuspects (NS) from Case Study 1

Label	Category	Subcategory	Words	Phrases	Patterns
SU	Manufacture	Poison	Cyanide		
SU	Buy	Fake prescription	Codeine, oxy, xans	Fake prescription	
SU	Sell		Drugs, opium, cannabis, salvia, marijuana	Leave immediately; liquid L; crystalized L	[careful AND sell]; [cover cops AND spot dealers]
NS			Steam, google, gmail, youtube		

human user, Fusion provides a user-friendly interface with the lexicon in a hierarchical form. Users can manage lexicons by viewing a history of changes and updating accordingly.

We represented the expert's conceptualization of classification rules as concept maps similar to decision trees (Vanides et al. 2005). In contrast to lexicons, concept maps represent the links between concepts and decisions independent of particular words in the text, allowing a richer representation of relationships between concepts (Novak 2010). Figure 3 shows a concept map in which nodes represent concepts that are classification criteria, and arcs represent optional paths leading directly or through other concepts to SU or NS. In this example, the concept map includes two concepts, namely technical expertise level and operational security, which the domain expert (DE2) chose to consider when classifying messages as SU or NS. The callout pointing at an arc (equivalent to a branch in a decision tree) explains the arc and its preceding arcs on the classification path taken. For example, a message is classified as SU if it is of high technical expertise AND exhibits operational security. In an accompanying table, the user explains the terms used in the concept map and the rationale for using them as classification criteria. At this stage, Fusion supports the generation of the map and table, but the data scientist uses it manually to redesign the ML classification model (corresponds to the feedback loop Human feedback for ML in Figure 1).

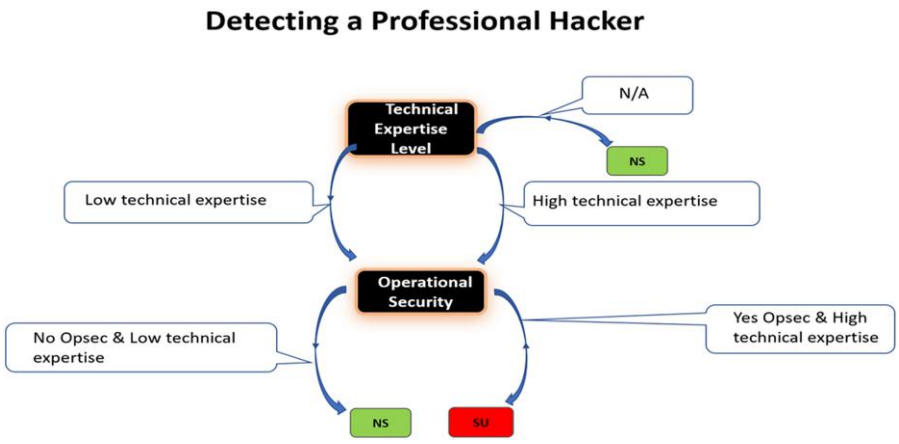
We guided the experts to use two suitable qualitative data analysis techniques (Silverman and Sommer 2019) to create concept maps and lexicons: The first is *qualitative content analysis*, which combines concept-driven and data-driven hierarchical categorization to guide classification (Riessman 2011, Schreier 2012). The resulting categories form part of the expert's conceptualization. The second technique is *secondary structural analysis* in which semiotics (words and other symbols) that signify the concepts are extracted from the categorized data in the previous phase (Grbich 2012). The qualitative content analysis considers, simultaneously, concepts and context, which is the perspective and background information leading to the concept (McTavish and Pirro 1990).

Using these qualitative analysis techniques iteratively, the expert generates a concept map and a lexicon of categories and related terms. The qualitative hierarchical categorization assists the expert in determining the concepts and the relationships between them, as seen through the expert's perspective. For example, the initial message "I want 50000 USD" in the Hidden Answers forum, was labeled Suspect because the expert understood the context as relating to illegal rather legal activities after reading the following reply to the message: "Go be a hitman or something, sell drugs on DREAM market." The visual display and interactive manipulation of the objects on the map help the user create the structure and group, avoid duplication of concepts, and explore alternatives (Bazeley and Jackson 2019).

With the structural analysis, the expert determines words (unigrams), phrases (n-grams) and patterns (skip-grams) that are associated with the concepts by similar meanings rather than frequencies of co-occurrence (Zhao and Mao 2018). These terms are organized hierarchically under concepts and subconcepts for the lexicon. The lexicons are fundamentally different from the well-established Linguistic Inquiry and Word Count (LIWC) lexicon (Tausczik and Pennebaker 2009). Instead of using a priori general categories as in LIWC (e.g., universal emotions), the domain expert generates a posteriori (i.e., data-driven) categories, and their ascribed semiotics are rooted in the specific corpus, as detailed previously.

The machine translates the lexicons generated by the expert as follows: first, the data scientist translates the lexicons to a bag-of-words (BoW) (e.g., the pattern [careful AND sell] will be added to the BoW as [careful_sell]). Next, each term in the BoW is replaced by its embedding vector. If the Smart-lexicon option is used, the lexicon is then enhanced via semantically similar

Figure 3. (Color online) Concept Map for BitsHacking Iteration 2 Created by DE2



terms, that is, nonlexicon terms for which the cosine similarity with at least one lexicon term is high (>0.9). Finally, for each observation, we compute the average embedding of the lexicon terms of which it is composed (a technique called "average pooling"). To translate the rules extracted by the domain expert, the data scientist generates features.

4.2.2. Supporting Multiple Perspectives in Reciprocal Learning (Emerging Design Principle 2). In RHML, contrasting human perspectives with alternative ML models promotes a dialogue that challenges or supports a perspective. Although perspective taking is a key mechanism for sense-making, it is often difficult to implement when working alone without training or computer-based support (Boland et al. 1994). Fusion guides the users in the process of offering their own perspective while taking in alternative ML perspectives. In selecting the combination of model type, lexicon, enhancement corpus, and concepts (Figure 2), the user controls the effect of human learning on the machine's feature engineering and thereby influences the outcome of the machine's classification process—that is, the machine's perspective, which is then shown to the user.

One of the difficulties in perspective taking is the complexity of examining a perspective that differs from your own, where each perspective involves ideas, relationships, and rationales that must be comprehended. We found that considering several alternative ML models proved to be too difficult at first. To ease perspective-taking, we designed a simplified interface (Figure 4), in which the user obtains feedback from only two ML models, namely Reflection and New Insights. The former helps users *reflect* on their current conceptualization as is, and the latter presents the machine's perspective to trigger new *insights* to expand their conceptualization. This simplified interface was used in the experiments described in Section 5.

4.2.3. Supporting Contextualization and Attention Management with Feedback (Emerging Design Principle 3). The learning cycles stimulated by the RHML configuration provide the opportunities for "listening and articulating" that are essential in reciprocal learning and furthermore the ability to manage the use of context and the attention to context through mutual feedback. Fusion supports the use of context in sense-making, both linguistic context and situational context, by drawing attention to parts of the context to be analyzed and presenting the context that explains texts under consideration. For instance, the human feedback to the machine reflects the attention given to context by the human, potentially guiding the use of context by machine. Similarly, the feedback to the human shows the machine's attention to context for the expert's consideration. In fact, the need for contextualization is most prominent in the design of the feedback, which must be given and understood in context, and is the primary vehicle for improving the learning process and its outcomes.

Fusion provides outcome feedback or explanatory feedback in accordance with the user's needs. Outcome feedback is provided in the form of performance metrics

Figure 4. (Color online) Simplified Interface with Only Two Classification Models (Reflection and New Insights)

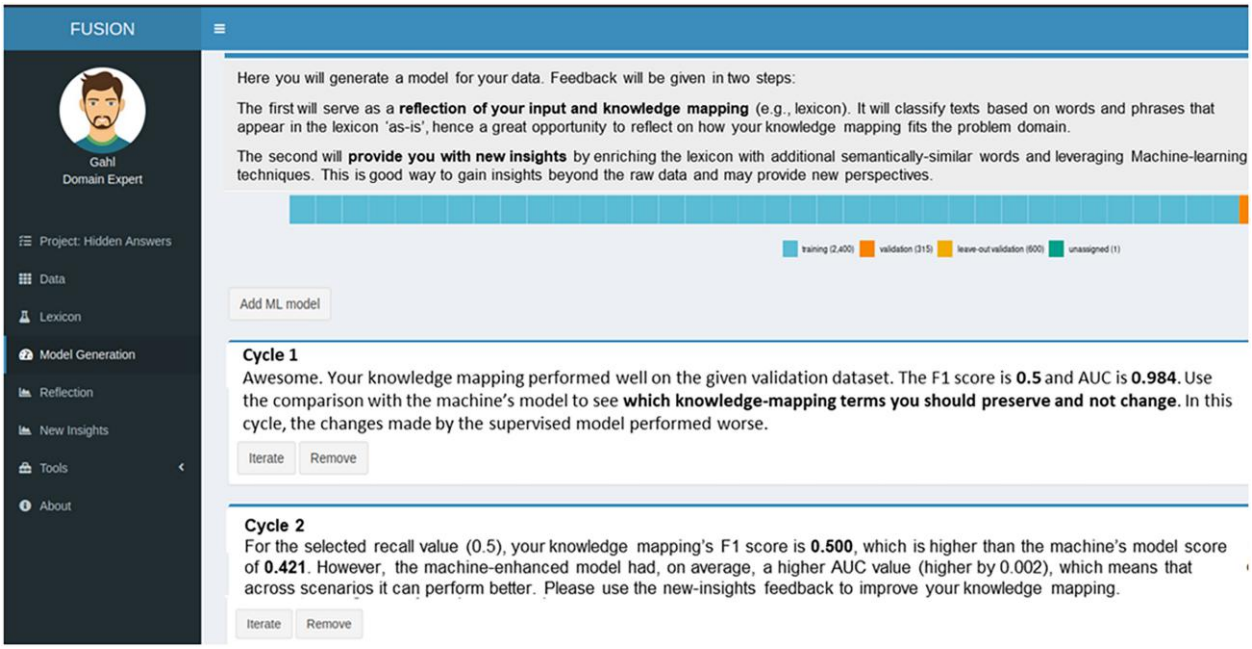
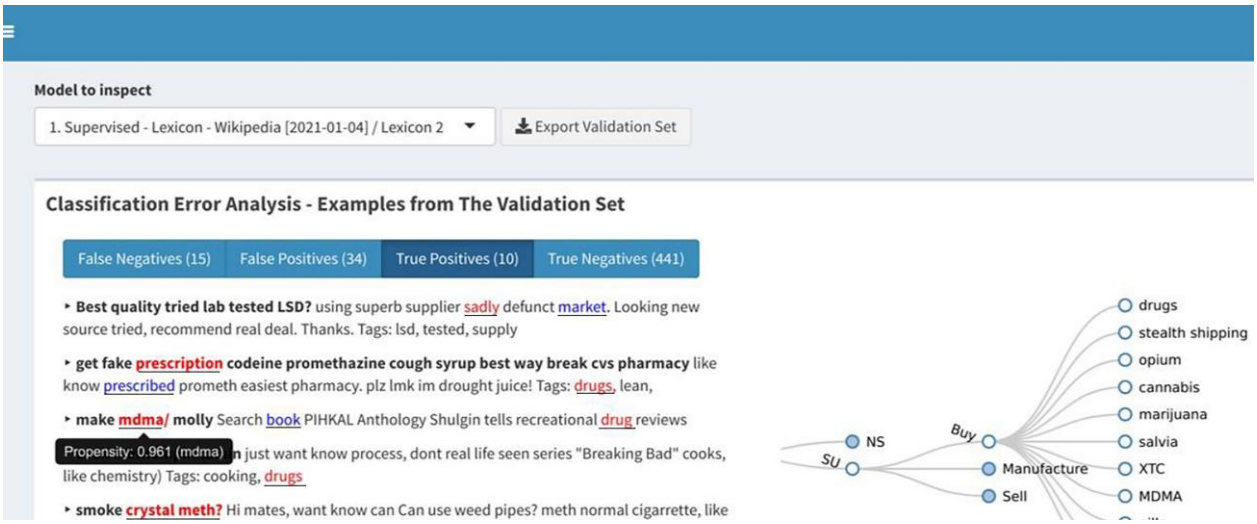


Figure 5. (Color online) Drill-Down Analysis of Concrete Text Messages



Note. On mouse hover, the user is shown the term propensity and the context of the word.

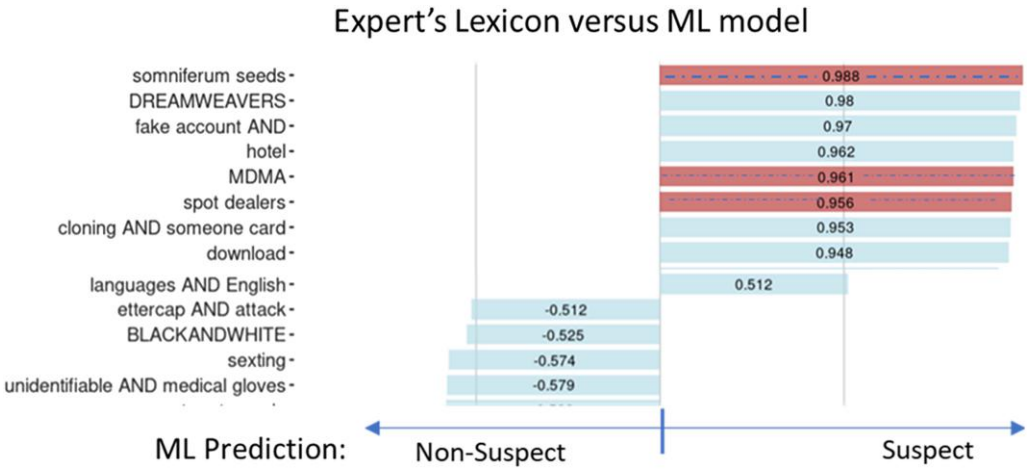
such as Area Under the Curve (AUC), recall, precision, accuracy, and a confusion matrix (lower part of Figure 2). Outcome feedback is displayed whenever learning is monitored, for example, when generating or comparing models. It is considered high-level feedback, often displayed as a bird's-eye view.

Explanatory feedback can be used in sense-making to "drill-down" on specific cases and terms to better understand decisions and performance. In Figure 5, classified messages are given in the context in which they were communicated and classified by the Supervised-Lexicon-Wikipedia model. The messages are organized according to the confusion matrix, and the 4 (of 10) messages displayed are those correctly identified as suspects (true positives). The red and blue underlined words are terms that increased the model's propensity to classify a message as suspect or nonsuspect. The marked word "mdma" is shown in its linguistic context within the message, that is, the words around it. The marked word can also be seen in its context across messages through another function. By hovering over the term, the user can see its propensity to classify the message and the other lexicon terms that affected this propensity. In addition, the interactive category-tree (depicted on the right side of Figure 5) shows the term located in the expert's lexicon. The tree allows the user to view the term in its situational context.

Fusion supports the user's attention management by mapping the differences between attention to features used by the ML model and attention to concepts by human experts represented in their lexicon. These differences are illustrated in Figure 6, which shows, for each term, the human expert's propensity to classify (suspect or nonsuspect) versus the ML model's propensity. The human propensity is binary, where broken line (and red) denotes suspect and blue denotes nonsuspect; the machine model propensity is calculated on a scale of −1 to +1, where positive numbers denote suspect and negative numbers denote nonsuspect, and the absolute value denotes low or high propensity. For example, the first term, "somniferum seeds," is marked with a broken line (and red) and has a high positive value, thus showing the ML fully conforms with the human assignment. In contrast, the second term "DREAMWEAVERS"[3] also has high positive value but is colored blue, which shows strong disagreement between the human and machine.

4.2.4. Providing Explainability of Black Box Models (Emerging Design Principle 4). Section 3 discussed how a human's capacity to interpret ML outcomes and to provide machines with relevant feedback is strongly dependent on algorithm explainability, that is, the ability to explain machine algorithms in ways understandable to humans (Lipton 2018, Carvalho et al. 2019, So 2020b). Indeed, explainability became a critical factor for Fusion users. Initially, we implemented the Local Interpretable Model-Agnostic Explanations (LIME6) software package as a routine component to support explainability (Ribeiro et al. 2016). As Fusion was increasingly used to analyze large data sets (>500 messages at a time), LIME6 became impractical, such that the interactivity we considered essential came at a cost of extremely slow processing. We therefore developed our own explainability components, which used the generated model to predict term propensity (for N-grams and skip-grams) and then distinguished the words (unigrams), phrases (N-grams), and patterns (skip-grams) that the model used for classification. This solution allowed us to control and customize user feedback. For example, we used

Figure 6. (Color online) How Terms Were Assigned by the Domain Expert (Suspect by Broken Line (and Red); Nonsuspect by Solid Marker (and Blue) vs. ML Model (Suspect by Positive Numbers, Nonsuspect by Negative)



these components to provide the feedback in Figure 6. In experimenting with explainability, we found that users experienced information overload when presented with an explanation of propensity calculations. To reach the right level of detail, we began by color-coding words (either in red or blue). Based on observations from the user studies, we gradually added relative propensity scores per word, which helped communicate how the model reached the final classification, and added context to the coloring, showing the phrases or patterns that included the colored word.

4.3. Use and Evaluate Fusion

We observed the use of Fusion primarily through structured user studies, carried out at the end of each design ("Build") iteration. In each user study, the domain experts (the users) were asked to classify messages from their respective corpora: Hidden Answers or BitsHacking.

To recap, a typical scenario of a *learning cycle* using Fusion can be summarized as five steps, where each step is performed either by the human or by the machine (their technical details were provided in the previous subsection). The five steps are (1) expert generates concept maps and derive lexicons and rules; (2) machine reads the lexicon and the rules, translates them into unigrams, n-grams, skip-grams, and features (rules), and then to embeddings; (3) machine runs classification models, both unsupervised (scoring, does not use embedding), and supervised ML (on the pooled embeddings); (4) machine returns outcome and explanatory feedback to expert (False Positives, False Negatives and feedback on the lexicon); and (5) experts see the feedback and, at their discretion, manually revise the concept maps and lexicons to reflect their knowledge more accurately.

4.3.1. Fusion Evaluation Procedure. We conducted user studies and cognitive walkthroughs to evaluate the functionality and usability of Fusion. The user studies were conducted as part of the two case studies based on the different corpora as described previously (Hidden Answers and BitsHacking). The studies included observations of the experts using Fusion for several hours in each iteration, followed by interviews. In these interviews the automatic logs of the users' choices were retrospectively reviewed. The users who participated in the studies were the three domain experts. Each expert was asked to classify text messages, study the machine feedback with alternative ML models, and adjust their knowledge representations. The experts worked by themselves at home, accessing Fusion remotely, and in a continuous monitored session in the presence of one or two observers. At home, the experts engaged in sense making that resulted in new concept maps and lexicons, which they uploaded to Fusion. During the monitored session, we observed the experts' behavior, encouraged them to think aloud, and recorded their talk and screen activities. Each such session was followed by a semi-structured interview around the classification task and the benefits of and difficulties in using Fusion.

4.3.2. General Insights and Specific Lessons Learned. Most of the lessons learned from the user studies were implemented in Fusion's interface (in successive Build iterations) and are discussed in the previous Build Fusion section. In what follows, we note several lessons on functionality and usability that we wish to underscore. Furthermore, we learned from the user studies that domain experts need time to reflect, re-examine, and rethink their conceptualizations, more than we expected.

Explainability: The most pressing observed difficulties were tied to the inadequate explainability of the feedback, both outcome and explanatory. DE1 wanted

to know which parts of the input data were considered by the ML models and wanted better explanations of the performance metrics provided by the models. DE1 was impressed that the interface offered the user the option to use "brushing" (i.e., selecting and filtering messages by painting sequences of messages on which to concentrate), an option that helped to focus on manageable sets of interest.

Use of Lexicon Terms: Another key challenge the domain experts faced was the dissonance between how they assigned a term in their lexicon and how the machine used it to classify. For example, DE1 saw that the supervised lexicon-based model gave some terms such as "Marijuana" a high propensity score for nonsuspect, although DE1 had marked this term as suspect in his lexicon. This confused DE1, as he wanted to know how this "mistake" had happened. DE2 also encountered this dissonance and described the propensity scores as a "nice to have feature"—explaining that unexpected propensity scores might trigger an examination of why the machine assigned these scores, either by examining all text messages that contained the specific term, or by using the term-frequency aggregated table.

The use of lexicons was revised in the second Build iteration. We added components that provide the necessary context for low-level analysis and improved cognitive support. Two examples of this were shown in Figure 5: the interactive category-tree defined by the user and the hover-over feature showing which lexicon-terms affected the propensity score, thus helping the user see the rationale for the classification choice.

Model Generation Controls: In our fourth Build iteration, we also provided a simplified model-generation screen for domain experts with little data science experience. In contrast to the multiple possible combinations shown in Figure 2, the simplified screen (Figure 4) offers only two options, namely, Reflection and New Insights. These options provide the results of two different ML classification algorithms. The Reflection algorithm corresponds to the case in which classification is based only on the user's conceptualization (it is a combination of an unsupervised model with a lexicon); feedback from this algorithm enables the user to observe the performance of his or her conceptualization "as is." The New Insights option, in turn, provides results of the supervised ML model with a smart lexicon using, in this case, Wikipedia (consistent with Devlin et al. 2018); this option enables users to view algorithmic classification suggestions that may influence their own conceptualizations. Additionally, we reduced the number of performance indicators on the screen to only two, namely AUC and F1 (F1 score is defined as the harmonic mean between precision and recall).

Use of Color: A general usability issue that we believe is especially important in human-machine learning systems is related to the color-coding of the words; we used red to indicate suspect words and blue to indicate nonsuspect words, but, initially, we did not explain this choice. Both DE1 and DE2 needed time to adjust to this color-coding and to interpret what each color represented. We addressed this issue in the second round of user studies. Interestingly, when asked to find improvements to the classification of messages, DE1 focused mainly on words that were colored, that is, words that already appeared in the lexicon, whereas DE2 first explored the whole text message, and then examined the colored words. We had to draw the expert's attention to the noncolored words and explain their potential value when updating the lexicon. The problem was alleviated when using only two models, namely, Reflection and New Insights (see Model Generation Controls).

5. Summative Evaluation of Reciprocal Learning: Two Experiments

Following development of the technology artifact that embodied the RHML theory, we conducted two experiments in which our domain experts and machine models engaged in multiple learning cycles. These evaluations were designed to assess the contribution of our final RHML configuration to both human and machine learning.

5.1. Data, Participants, and Procedure

For these experiments, we constructed two new corpora from the Darknet forum data described in Section 4. The participants in this study were the same domain experts who had participated in our user studies; each was assigned to the corpus corresponding to their expertise (DE1: Hidden Answers; DE2 and DE3: BitsHacking). Each corpus contained a set of 3,000 observations, for which external domain experts had generated ground-truth labels. Each corpus was split into nine equal-sized batches. Then, the participant assigned to the corpus used Fusion to analyze the corpus over eight learning cycles, with one batch being analyzed per cycle. The last batch was used as a leave-out-set for evaluation purposes. In each learning cycle, the participant was asked to classify text messages, as well as to provide a concept map and lexicon. In turn, the current cycle's batch (used as a training set) and the next batch (used as a validation set) were classified by three ML algorithms—Baseline, Reflection, and New Insights, as described in Section 4. In cycles 2–8, the participant reviewed feedback on the validation set from the Reflection and the New Insights ML models (Figure 4) and subsequently adjusted their knowledge representations (lexicon and concept map) when needed, mainly by adding, revising, or deleting concepts and

relationships. At the end of each learning cycle, the participants responded to a survey in which they were asked to summarize their learning activities and perceptions.

5.2. Results and Discussion of Experiments

Table 4 summarizes the outcomes of the learning cycles for both the human participants and the three ML algorithms. To quantify *human learning*, we focused on the knowledge representations that participants generated. For the concept maps, we measured (i) the number of concepts in the map at the end of the learning cycle, and (ii) the relationships between these concepts (Novak 2010). Figure 7 illustrates the evolution of the concept maps for the BitsHacking case. For example, in the first map (corresponding to learning cycle 1), the domain expert identified two concepts, connected by five relationships. For the lexicons that participants produced, we used the idea of weighted attributes (Mokryn and Ben-Shoshan 2021) to measure the progression between lexicons as follows: (i) the absolute difference between the size of the current lexicon and that of the previous learning cycle and (ii) the weighted average of lexicon size: $\beta_1 \times terms + \beta_2 \times phrases + \beta_3 \times patterns$, where phrases are forms of n-grams, and patterns are forms of skip-grams. We set $\beta_1 = 1$, $\beta_2 = 2$, and $\beta_3 = 3$, in accordance with the average number of terms, n-grams and skip-grams in the lexicons.

To quantify *machine learning*, we used the AUC on the leave-out set. AUC is considered an appropriate measure for comparing classifiers' performance, as it does not depend on a cutoff value. It is important to note that the experts did not observe the algorithms' performance on the leave-out set, but rather the performance on the validation set. The results reported here provide an indication of machine learning as reflected in classification performance on an out-of-sample, unseen, set, even in cases where the expert or the machine over fit the learning to the observed data set.

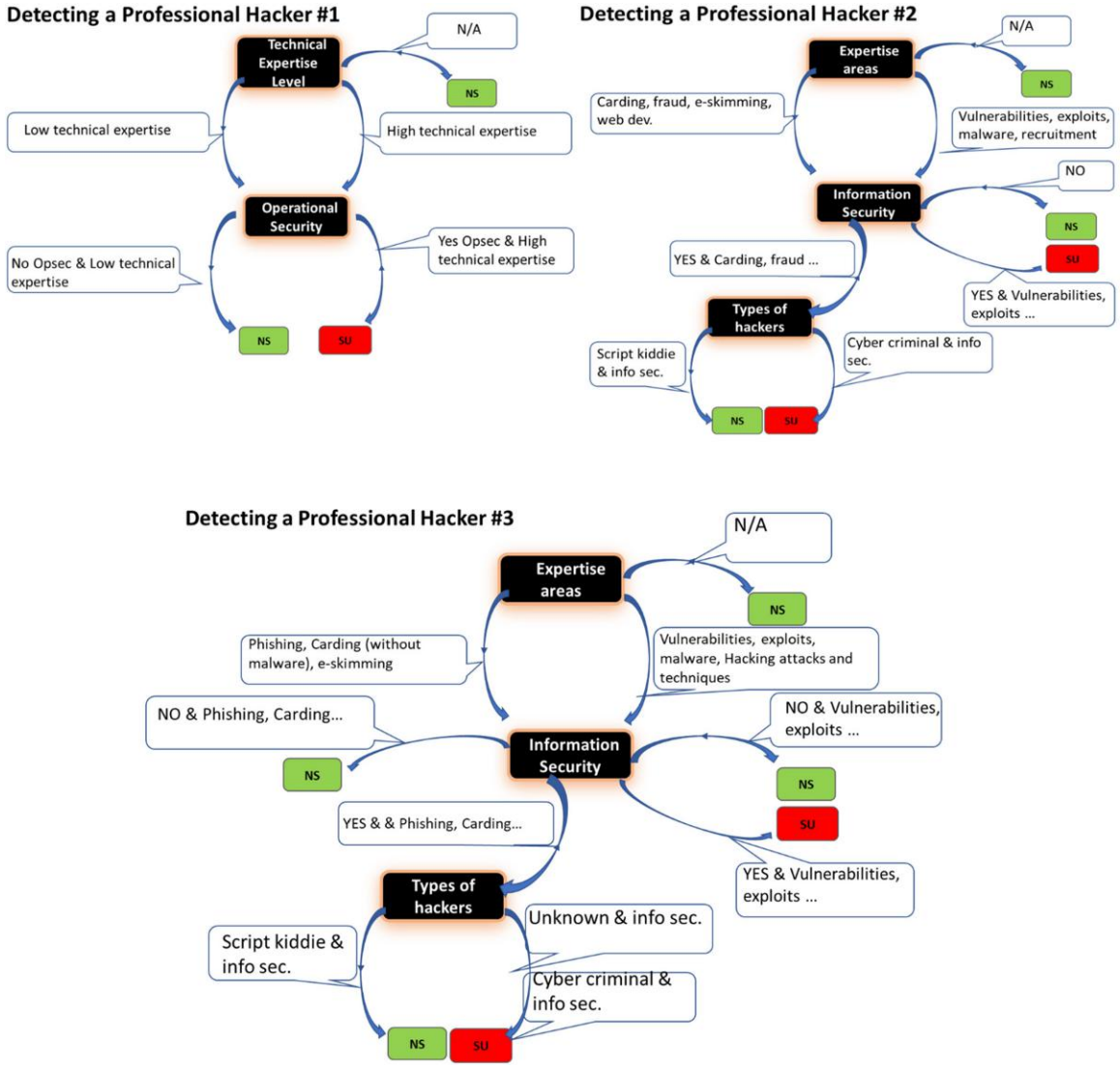
Observing the results, we can see the reciprocal learning of the two parties. In the Hidden Answers case, we observe that in the first three cycles, the expert's knowledge dominated the machine knowledge, as indicated by the higher performance of the Reflection algorithm compared with both the New Insights and Baseline algorithms. In the fourth learning cycle, the domain expert added a concept that differentiates an illicit drug transaction from a general discussion on drugs (Table 4). In essence, this concept better defined the boundary between the suspect and nonsuspect classes. This definition enabled the machine to enhance its learning, as can be observed by the dominant performance of the New Insights algorithm from the fifth learning cycle onward. This finding is in line with the claim that a broader context facilitates more effective processing of the data (Khashabi et al. 2020).

Regarding HL, we observe that the expert added knowledge to their concept maps and lexicons. Although this added knowledge did not explicitly translate into better over time performance in the Reflection algorithm, it did result in improved joint knowledge of the human and the machine, which translated into

Table 4. Summary of Human Learning and ML Measurements for Each Learning Cycle

		Human learning				Machine learning		
		Concept maps		Lexicon		Baseline	Reflection algorithm	New insights algorithm
Data set	Learning cycle	Number of concepts	Number of relationships between concepts	Weighted size of lexicon	Delta lexicon size	AUC	AUC	AUC
Hidden answers	1	6	10	448	NA	0.73	0.88	0.79
	2	6	10	485	29	0.58	0.88	0.64
	3	6	10	508	19	0.80	0.89	0.84
	4	7	12	510	1	0.83	0.79	0.70
	5	7	12	516	6	0.81	0.79	0.81
	6	7	12	549	23	0.77	0.79	0.84
	7	7	12	571	21	0.78	0.78	0.89
	8	7	13	650	59	0.75	0.79	0.87
BitsHacking	1	2	5	404	NA	0.67	0.82	0.77
	2	3	10	358	−32	0.59	0.72	0.61
	3	3	10	343	−13	0.55	0.78	0.50
	4	3	10	156	−107	0.59	0.69	0.67
	5	3	10	165	8	0.39	0.70	0.59
	6	3	10	190	21	0.58	0.77	0.63
	7	4	12	222	13	0.54	0.67	0.63
	8	4	12	224	1	0.50	0.67	0.66

Figure 7. (Color online) Evolution of Concepts Maps in the BitsHacking Case



better performance for the New Insights algorithm. Table 5 further presents selected comments that participants provided at the end of each learning cycle, indicating what they changed in their knowledge representations to move the learning process forward and which model they learned from.

Figure 8 summarizes the learning of the human and the machine across learning cycles. In the BitsHacking case, we observe a different behavior and learning pattern (right side of Figure 8). First, it is important that the average performance of the Baseline model on the leave-out set declined over time. This finding suggests that the data were nonstationary: Suspects in new data (in each batch) were plausibly different from those appearing in the leave-out set. The Baseline algorithm failed to capture the difference. The expert, on the other hand, was better at learning the concepts that were consistent over the different data batches. Although the expert also struggled at first to learn the data (as seen in the declining performance of the Reflection algorithm in the first three learning cycles), he managed to better represent the changing environment in the fourth cycle and stabilize his performance (note that he could not reach the performance of the first learning cycle, as the data changed). He did so by completely discarding his lexicon and constructing a new one (Table 5). The New Insights algorithm, which represents the ML, lagged behind the expert's performance, yet still outperformed the noninformed Baseline model. As elaborated in our discussion, this type of learning can support the significant challenge of ML degradation over time caused by domain-driven shifts in data characteristics.

Table 5. Comments from Experts After Each Learning Cycle, Indicating Key Actions Taken

Learning cycle	Hidden answers	BitsHacking
1	Initial development of concept map and lexicon	Initial development of concept map and lexicon
2	Added one category and enhanced lexicon	Merged subcategories, removed some lexicon terms according to the machine feedback
3	Grouped four subcategories into one due to thematic proximity	Removed additional terms that result in poor performance
4	Added concept to differentiate drug market from discussion of drugs; developed lexicon	Replaced the entire lexicon to better reflect the data
5	Refined lexicon according to the machine feedback	Refined lexicon according to the machine feedback
6	Improved the lexicon by adding two new patterns	Added professional terms detected by Fusion and generalized them in the revised lexicon
7	Refined lexicon according to the machine feedback	Removed terms according to Fusion's recommendation, and added new terms to clarify
8	Refined lexicon according to the machine feedback	Refined lexicon according to the machine feedback

6. Discussion

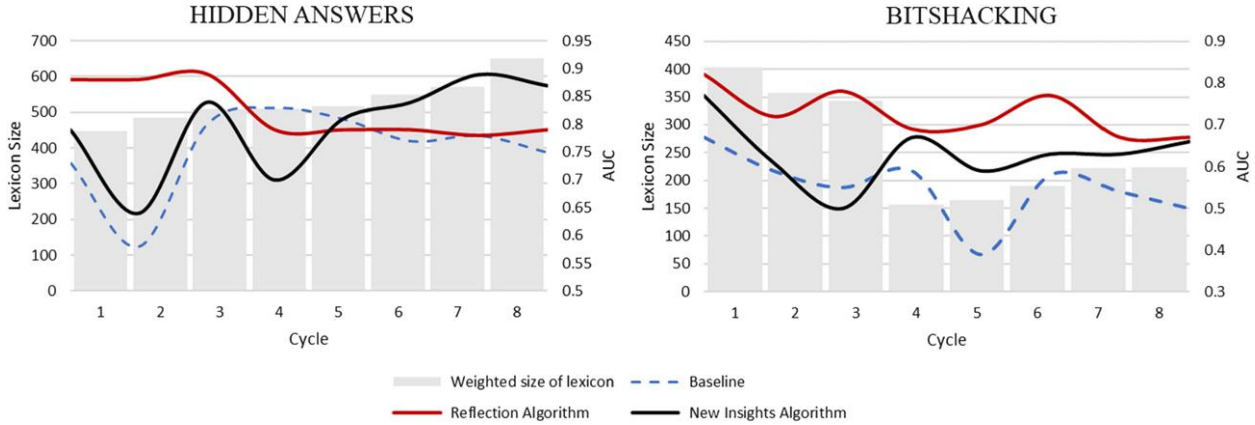
6.1. Summary and Knowledge Contribution

This work started out with a vision of human-machine collaboration that keeps the human in the learning loop (So 2020a). Following a DSR approach that integrated knowledge from multiple disciplines, we developed three artifacts that, collectively, achieve this vision. In line with conventions of DSR (Gregor and Hevner 2013, Lee et al. 2015, Schwartz and Yahav 2021), we articulate the knowledge contributions of each artifact according to the Knowledge Contribution Diagram explained in the appendix and discuss each artifact's contribution and limitations.

6.1.1. Theory Artifact: RHML Configuration (Exaptation). Our first knowledge contribution is the creation of a theory artifact, an abstract RHML configuration (Figure 1). In developing this artifact, we built on theory and practical insights from the domain of reciprocal learning in human dyads (Jörg 2004, 2009; Holzer and Kent 2013) and extended these principles to the domain of human-machine dyads. Specifically, we began with the premise that dialogue between human learning partners (Havruta) leads to shared meaning and sense-making through three mechanisms: (i) listening/articulating; (ii) wondering/focusing attention to establish context; and (iii) challenging/supporting perspectives (Kent 2010). We suggested that, in an RHML configuration, the various activities associated with learning could be allocated between human and machine to capitalize on their relative advantages, in a manner that ensures that the feedback from machine to human complements human processing abilities (and vice versa). We characterized our final configuration along three components: HL and ML processes, coordination, and control of learning activities through feedback and joint representations of classification knowledge. Feedback loops between HL and ML are the centerpiece of the RHML configuration, promoting cycles of learning and acting as a self-reinforcing mechanism (Jörg 2004, 2009).

Although prior work has introduced conceptual frameworks to describe collaborations between humans and machines—addressing specific considerations such as organizing, control, responsibility, and work

Figure 8. (Color online) Summary of Learning in the Two Cases



implications (Suchman 2007)—these frameworks were *macro-level* representations and left "collaboration" as an abstract concept lacking theoretical underpinnings. Our work, in contrast, builds a theoretical foundation for joint learning based on dyadic human learning, and then takes a *micro view* on the specific problem of RHML, explicitly addressing how the configuration operates, that is, the allocation of activities and the communication within the configuration—a necessary level of analysis for making these configurations operational. Given that human dyadic learning theories are well established, dating back hundreds of years, one can consider their application to the relatively new problem of human-machine dyads to be an *exaptation* (Gregor and Hevner 2013).

The RHML configuration is useful only for systems in which it is important to keep the human in the learning loop. We concentrated on learning and sensemaking but believe our conclusions also apply to other forms of intelligent collaboration, such as joint decision making and design, going beyond the domain of cybersecurity and beyond the specific task of text-message classification. For instance, screening medical images is a well-established mechanism used globally for early detection of diseases, but because of the vast number of images and insufficient trained experts, there is a lack of manpower to handle the increasing workloads (Budd et al. 2021). AI-based screening systems have been shown effective in achieving higher accuracy in interpreting images, often as high, or even higher, than trained experts, such as in the example of detecting breast cancer from mammograms (McKinney et al. 2020). Nevertheless, in other cases, human detection outperforms the machine detection, as human readers possess a broader context of the patient's history when making decisions. This highlights the need for machines to learn from humans (Budd et al. 2021). RHML configurations would enable periodic learning sessions in which medical experts and AI systems update their diagnostic knowledge. Yet another, more remote domain in which RHML can be applied is the exciting area of collaboration between experts and robotics in industry, where the experts must continue to learn to guide and control the robots (Ansari et al. 2018).

6.1.2. Technology Artifact: Fusion (Improvement). Our second knowledge contribution stems from the creation of Fusion, which serves as an instantiation of our abstract RHML configuration. We developed Fusion in four Use-Build-Evaluate iterations, in which human experts engaged with the system to classify text from two Darknet forums: Hidden Answers and BitsHacking. We subsequently carried out two experiments, each involving eight learning cycles, to evaluate the system's capacity to support human and machine learning. The results of the latter experiments, discussed in depth in Section 5, provide clear evidence of the mutual-learning processes that both human and machine underwent. Notably, when we compared the classification performance of ML algorithms that incorporated different elements of the reciprocal learning process—the baseline with no human input (ML-only approach) versus classification based only on expert conceptualizations (Reflection) versus fully reciprocal learning (New Insights)—we observed nonlinear performance trends over the eight learning cycles. For instance, in the BitsHacking case, the Reflection algorithm outperformed both the New Insights algorithm and the baseline algorithm for several learning cycles. Such a pattern is to be expected, as the impact of learning evolves in multiple dimensions over time. For example, when team members develop a new joint decision strategy, their accuracy initially declines and only later improves (Sengupta and Te'eni 1993).

Evaluating learning of text classification presents several challenges. In particular, we had no external ground truth but rather so called "weak classification" by expert labeling. This issue can be mitigated by incorporating multiple experts, as we did in our second case study. We allowed the experts to go back and evaluate their initial classification in light of the new learning, and we observed very few cases (less than 0.1%) where the experts changed their classification.

The design, development, and testing of Fusion addressed the well-established problem of text classification with a novel approach to both interface and interaction. We suggest that, by explicitly supporting learning cycles between human and machine, Fusion represents a clear *improvement* (Gregor and Hevner 2013) over existing HITL approaches aimed at text classification.

Fusion's architecture instantiates the RHML configuration of Figure 1 that incorporates classification algorithms stored externally and provides the functionality described in the menu of Figure 2. This design can be extended to other types of classification by incorporating new classification algorithms to include, say, image classification. For instance, currently we use binary classification algorithms, but they can be easily replaced with algorithms for multiclass or multilabel classification (such as emotion detection). Moreover, the system can also be extended to support other forms of collaborative classification. For instance, there is also potential learning that occurs between multiple human agents interacting with an RHML system, such as a domain expert learning to communicate with a data scientist. The current version of Fusion supports interaction of a single expert with a machine model, and this limitation does not allow the reconciliation of multiple expert views. Nevertheless, the RHML configuration can be augmented to support multiple experts, and we expect that further development of

Table 6. Derived Design Principles for Developing RHML Support Systems

Design principle	Explanation	Operationalization in Fusion
1. System should facilitate joint knowledge representations that develop through cyclic, continual and accumulative reciprocal learning	RHML configuration requires learning by machine and human. Supporting human learning requires learning cycles with limited new information at each stage; learning should continue as long as the human is in the learning loop. Similarly, machine learning should be iterative and accumulative.	Functional (a) Knowledge accumulates in a visible conceptualization; (b) Incremental learning adds small chunks of information per cycle; (c) Human-computer interaction feedback directs and restricts action; (d) Accessible logs of progress across cycles; (e) Ability to navigate across cycles. Communication (f) Conceptualization presented to fit the human expert's mental model.
2. System should support perspective taking	Taking and examining alternative perspectives is essential for effective conceptualization. In RHML, each ML model generates a perspective that is communicated through the machine feedback to the human with its unique added value. Perspective taking encourages exploration of new contexts through a new lens.	Functional (a) Adding new model types throughout the learning process, subject to rules for avoiding overfitting and the human's capacity to process. This functionality enables differential learning in supervised vs. unsupervised models; (b) Highlighting differences between elements of the perspectives. Communication (c) Feedback on alternative models should fit the user's mental model.
3. System should support contextualization and attention management with bidirectional feedback	RHML is supported with bidirectional feedback that explains, corrects and shifts attention to alternative and changing contexts. Feedback to the human: Explanatory feedback guides and explains classification, and outcome feedback assesses the individual's own learning. Feedback to the machine based on the expert's revised knowledge (a) points the machine to promising parts of context for improved ML, and (b) expands the linguistic context of words.	Functional (a) Outcome (high level) feedback in the form of classification accuracy measures; (b) Explanatory (specific) feedback at the message level; (c) Explanatory feedback organized by alternative dimensions (e.g., false vs. true classifications). Communication (d) Feedback to human indicates the mapping between human and machine conceptualization; (e) Feedback to human indicates how a change in human conceptualization is modeled by the machine.
4. System should provide explainability in feedback from machine to human	Explainability ensures effective communication so that the human expert understands the feedback and the reasoning behind it for effective learning.	Communication (a) Coloring the words in a message that impact the classification; (b) showing their propensity toward the assigned classification.

the technology artifact will lead to this important extension. We envision an environment in which not only will human and machine learn reciprocally, but so too will multiple humans interacting with the same machine.

6.1.3. Design-Principles Artifact (Invention). Our third knowledge contribution relates to the insights we gained over the course of the design process (Section 4), specifically, insights pertaining to the ongoing integration of a human agent into the ML process. (We devoted less discussion to challenges we faced regarding the algorithmic design for ML, as these types of challenges have received more research attention; So 2020a). We identified four main requirements that an RHML configuration must support: (i) joint knowledge representations; (ii) multiple perspectives; (iii) contextualization and attention management through feedback; and (iv) explainability of black box ML algorithms. We formulate our insights as a set of design principles in Table 6. The table explains each principle and its operationalization in Fusion at (i) the functional level—which refers to functionality provided by human or by machine to fulfill the requirement; and (ii) the communication level—referring to how information should be presented.

As Fusion is an evolving research environment, new functionality will be developed to fully realize the four design principles. Correspondingly, as we gain more experience and expand Fusion's support for diverse

learning activities, the set of design principles is likely to expand. For example, future additions may relate to the problem of transferring control from human to machine.

We emphasize that the overarching design goal for all four principles is to facilitate effective joint learning of both human and machine over time. Though different problem domains may introduce different requirements for such learning, we suggest that the principles we propose are likely to transcend specific contexts and to serve as a general basis for the development of RHML configurations. Given that, to our knowledge, no formal design principles exist for creating RHML configurations, the knowledge contribution of our design-principles artifact falls in the invention quadrant of Gregor and Hevner (2013).

Design principles for a new type of software system depend on several factors, including the purpose of the system, the intended user base, and the technology used to build the system. Although there may be overlap between the RHML design principles and others in common use, the discernment of which design principles are most relevant, and how they should be applied, represent a significant new insight unique to RHML systems in that the principles are applicable to both human and machine agents.

6.2. Limitations and Future Work

Our work has several limitations, which may motivate future work. First, our case studies involved corpora from the Darknet. Open communication among known friends may introduce new challenges that necessitate further adaptation of our RHML configuration and interface. Moreover, if we regard RHML as a form of responsive design (Germonprez et al. 2017), in which the machine and the human continue to redesign their actions by learning from each other, we should explicitly consider social and emotional aspects of reciprocal learning (Holzer and Kent 2013). Second, our current configuration requires a knowledgeable expert classifier; a requirement that introduces its own limitations, as experts' judgment is subjective, which threatens the validity of their decisions. Indeed, research on augmented text classification has identified concerns related to duplication of human bias in ML algorithms and low interrater reliability when establishing ground truth (Duarte et al. 2018). In our studies, we took steps to mitigate expert bias and inconsistency to the extent possible by instructing study participants to use recommended methods of qualitative analysis (Creswell and Creswell 2017) in creating their conceptualizations of classification knowledge. Moreover, for the BitsHacking corpus, we added a second external expert to increase reliability and minimize bias. Nevertheless, we are seeking ways to follow up on a sample of suspect messages to compare expert predictions against cases that were subsequently identified, establishing objective ground truth. Expert bias highlights a potential benefit of the RHML configuration: Such a system can enable both human and machine to point out bias through their mutual feedback.

We further acknowledge that, in selecting our ML algorithms, we did not carry out an exhaustive search for the most suitable approaches (bag-of-words, word2vec). We are currently exploring the use of alternative algorithms and developing new forms of conceptualizations that might enhance communication between humans and machines, as well as algorithm explainability.

On a related note, our work emphasized the importance of joint knowledge representations that support communication between human and machine while enabling both parties to fulfill their respective functions. In Fusion, these representations currently include hierarchical lexicons of concepts (categories) and concept maps. We acknowledge that these representations might not capture all new knowledge gained from the human expert's learning. Future research should explore the development of alternative mappings and the automatic translation of experts' maps into machine readable representations (e.g., knowledge graphs and ontologies; Ji et al. 2021). More broadly, there remains a design challenge in providing functionality for a generative memory, much like the multiple modes of information processing available in human memory (Goldstein 2014). We believe that visualization of knowledge representations will also become a significant area of research for RHML configurations (So 2020b).

6.3. Broader Implications for the Role of the HITL

Beyond the practical considerations associated with the goals and operational challenges of keeping humans in the loop, prior research has raised more abstract questions regarding how human-centered approaches affect machine learning and vice versa (Gillies et al. 2016, So 2020a). These questions can be formulated in our research context as follows. (1) What role do humans play in RHML configurations? (2) Do RHML configurations change the ways in which human learning and machine learning are done, and, if so, how? We believe that our work helps answer the first question and takes a first step toward answering the second question. Our evaluation methods encompass both machine learning and human learning showing improvements in the overall RHML configuration. In keeping with the Gestalt model of DSR (Adam et al. 2021), we have advanced a prescriptive design theory for achieving synergies among humans and machines. By defining and evaluating the allocation of learning activities within an RHML configuration, we can better understand and define the role of human experts who learn

from artificial intelligence, and, at the same time, provide insights that can contribute to machine learning. Specific allocations of learning activities will need to be reexamined in light of new AI developments.

Our vision is for the RHML configuration to serve as a theoretical framing for any interaction in which both human and machine can learn from each other. We believe that RHML can be applied in many scenarios where both humans and machines can learn together by doing, and where human development is valued.

Acknowledgments

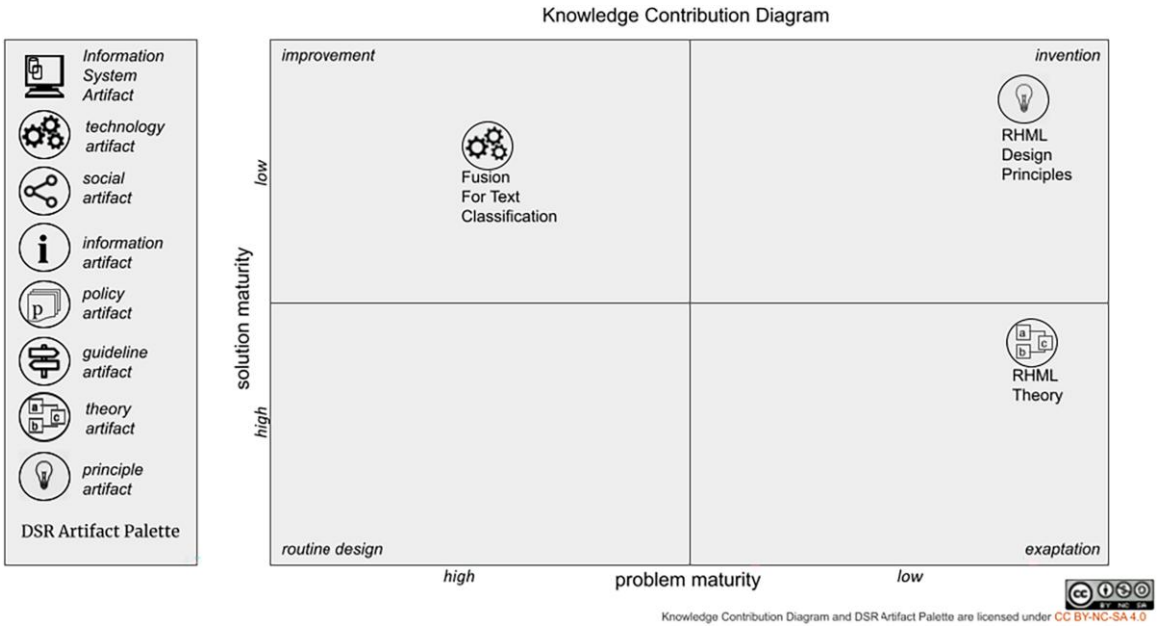
The first two authors contributed equally to this work. This study benefitted from the generous support of the Jeremy Coller Foundation and the Henry Crown Institute of Business Research in Israel.

Appendix. Knowledge Contribution Diagram

A knowledge contribution diagram (KCD), based on the knowledge contribution framework of Gregor and Hevner (2013), is a graphical technique used to present the knowledge contribution of complex design science research projects, especially those that involve multiple artifacts with multiple knowledge contributions of distinct types (Schwartz and Yahav 2021). The KCD can help researchers and designers communicate their design science research more effectively and expressively by concisely visualizing their artifacts and contributions. Using the axes of problem maturity and solution maturity, each artifact is considered individually, and positioned as follows (see Figure A.1):

- RHML, a theory artifact, represents an exaptation of Jorg's dyadic human learning theory into the realm of human-machine learning. Although the existing theories of dyadic human learning are relatively mature, their application to human-machine dyads in a new problem space are not, justifying placement of this artifact in the lower area of the *exaptation* quadrant bordering on the *invention* quadrant.
- Fusion, a technology artifact for text classification used to demonstrate and test RHML, represents an improvement on prior text classification systems by presenting a new solution to a known problem. Text classification is a mature problem addressed by many different solutions. The improvement shown through RHML in our experimental results justifies placement of this artifact in the center of the *improvement* quadrant.
- RHML DP, a design principles artifact, embodies a novel contribution to the immature problem space of reciprocal human-machine learning where design principles have yet to be considered, justifying placement in the upper right of the *invention* quadrant.

Figure A.1. Knowledge Contributions of Three Distinct Artifacts



Endnotes

[1] To avoid confusion: the concept of a "DSR iteration"—used to describe an iteration in our design process—is distinct from the concept of a "learning cycle," which is a cyclical process that takes place during the reciprocal learning interaction between human and machine.

[2] Implementation of Fusion is available at https://github.com/TAUCollerLab/Fusion.

[3] "Dreamweaver" is a term that is often used to describe a drug dealer. In our training data set, however, this term did not appear in the context of drugs, and therefore, the domain expert did not initially include it.

References

Abbasi A, Chen H (2008) CyberGate: A design framework and system for text analysis of computer-mediated communication. *Management Inform. Systems Quart.* 32(4):811–837.

Abbasi A, Zhou Y, Deng S, Zhang P (2018) Text analytics to support sense-making in social media: A language-action perspective. *Management Inform. Systems Quart.* 42(2):427–464.

Abdel-Karim R, Reda Y, Abdel-Fattah A (2020) Nanostructured materials-based nanosensors. *J. Electrochemical Soc.* 167(3):037554.

Adam MT, Gregor S, Hevner A, Morana S (2021) Transactions on Human-computer interaction. *AIS Trans. Human-Comput. Interactions* 13(1):1–11.

Amir O, Doshi-Velez F, Sarne D (2019) Summarizing agent strategies. *Autonomic Agent Multi Agent Systems* 33(5):628–644.

Ansari F, Erol S, Sihn W (2018) Rethinking human-machine learning in industry 4.0: How does the paradigm shift treat the role of human learning? *Procedia Manufacturing* 23:117–122.

Baird A, Maruping LM (2021) The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *Management Inform. Systems Quart.* 45(1):315–341.

Baralou E, Tsoukas H (2015) How is new organizational knowledge created in a virtual context? An ethnographic study. *Organ. Stud.* 36(5):593–620.

Bazeley P, Jackson K (2019) *Qualitative Data Analysis with NVivo* (SAGE Publications, London).

Boland RJ, Tenkasi RV, Te'eni D (1994) Designing information technology to support distributed cognition. *Organ. Sci.* 5(3):456–475.

Budd S, Robinson EC, Kainz B (2021) A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Anal.* 71:102062.

Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: A survey on methods and metrics. *Electronics (Basel)* 8(8):832.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2021) shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny.

Creswell J, Creswell J (2017) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (Sage Publications, London).

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint, submitted October 11, https://arxiv.org/abs/1810.04805.

Diedrich A, Eriksson-Zetterquist U, Styhre A (2011) Sorting people out: The uses of one-dimensional classificatory schemes in a multi-dimensional world. *Cultural Organ.* 17(4):271–292.

Duarte N, Llanso E, Loup AC (2018) Mixed messages? The limits of automated social media content analysis. *Proc. 1st Conf. on Fairness, Accountability and Transparency* (Center for Democracy and Technology, Washington, DC), 106.

Enarsson T, Enqvist L, Naarttijärvi M (2021) Approaching the human in the loop: Legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Inform. Comm. Tech. Law* 31(1):123–153.

Fitts PM (1951) *Human Engineering for an Effective Air-Navigation and Traffic-Control System* (NRC Committee on Aviation Psychology).

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* 33(1):1.

Fromkin V, Rodman R, Hyams N (2011) *An Introduction to Language*, 9th ed. (Wadsworth, Boston).

Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Inform. Systems Quart.* 45(3):1527–1556.

Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu N, Peters M, et al. (2018) Allennlp: A deep semantic natural language processing platform. Preprint, submitted May 31, https://arxiv.org/abs/1803.07640.

Germonprez M, Kendall JE, Kendall KE, Mathiassen L, Young B, Warner B (2017) A theory of responsive design: A field study of corporate engagement with open source communities. *Inform. Systems Res.* 28(1):64–83.

Gillies M, Fiebrink R, Tanaka A, Garcia J, Bevilacqua F, Heloir A, Nunnari F, et al. (2016) Human-centered machine learning. *Proc. CHI Conf. Extended Abstracts on Human Factors in Comput. Systems* (ACM, New York), 3558–3565.

Goldstein EB (2014) *Cognitive Psychology: Connecting Mind, Research and Everyday Experience* (Nelson Education, Ontario).

Grbich C (2012) *Qualitative Data Analysis: An Introduction* (Sage, London).

Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. *Management Inform. Systems Quart.* 37(2):337–355.

Groh M (2022) Identifying the context shift between test benchmarks and production data. Preprint, submitted September 22, https://arxiv.org/abs/2207.01059.

Grønsund T, Aanestad M (2020) Augmenting the algorithm: Emerging human-in-the-loop work configurations. *J. Strategic Inform. Systems* 29(2):101614.

Hevner A, March S, Park J, Ram S (2004) Design science research in information systems. *Management Inform. Systems Quart.* 28(1):75–105.

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

Holzinger A, Weippl E, Tjoa AM, Kieseberg P (2021) Digital transformation for sustainable development goals (SDGs) – A security, safety and privacy perspective on AI. Nugent R, ed. *Proc. Internat. Cross-Domain Conf. for Machine Learn. and Knowledge Extraction*, Part of Lecture Notes in Computer Science (Springer, Cham, Switzerland), 1–20.

Iivari J (2015) Distinguishing and contrasting two strategies for design science research. *Eur. J. Inform. Systems* 24(1):107–115.

Ip W, Damodaran L, Olphert CW, Maguire MC (1990) The use of task allocation charts in system design: A critical appraisal. Diaper D, Gilmore DJ, eds. *Proc. IFIP TC13 3rd Internat. Conf. on Human-Computer Interaction* (North-Holland Publishing, Amsterdam), 289–294.

Ji S, Pan S, Cambria E, Marttinen P, Philip SY (2021) A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Systems* 33(2):494–514.

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* 6(1):22.

Katz A, Te'eni D (2007) The contingent impact of contextualization on computer-mediated collaboration. *Organ. Sci.* 18(2):261–279.

Kent O (2010) A theory of Havruta learning. *J. Jewish Ed.* 76(3):215–245.

Khashabi D, Azer ES, Khot T, Sabharwal A, Roth D (2020) On the possibilities and limitations of multi-hop reasoning under linguistic imperfections. Preprint, submitted May 1, https://arxiv.org/abs/1901.02522.

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *Management Inform. Systems Quart.* 45(3):1501–1526.

Lee AS, Thomas M, Baskerville RL (2015) Going back to basics in design science: From the information technology artifact to the information systems artifact. *Inform. Systems J.* 25(1):5–21.

Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57.

Marcellino W, Johnson C, Posard MN, Helmus TC (2020) *Foreign Interference in the 2020 Election: Tools for Detecting Online Election Interference* (RAND Corporation, Santa Monica, CA).

McKinney SM, Sieniek M, Godbole V, Godwin J, et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577(7788):89–94.

McTavish DG, Pirro EB (1990) Contextual content analysis. *Qual. Quant.* 24(3):245–265.

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. Preprint, submitted September 7, https://arxiv.org/abs/1301.3781.

Mokryn O, Ben-Shoshan H (2021) Domain-based latent personal analysis and its use for impersonation detection in social media. *User Modeling User-Adaptation Interaction* 31(4):785–828.

Moreno R (2004) Decreasing cognitive load for novice students: Effects of explanatory vs. corrective feedback in discovery-based multimedia. *Instrument Sci.* 32(1):99–113.

Nonaka I, Takeuchi H (1995) *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation* (Oxford University Press, Oxford, UK).

Nonaka I, Toyama R, Nagata A (2000) A firm as a knowledge-creating entity: A new perspective on the theory of the firm. *Industrial Corporate Change* 9(1):1–20.

Novak JD (2010) *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations* (Routledge, London).

Nunamaker JF Jr, Twyman NW, Giboney JS, Briggs RO (2017) Creating high-value real-world impact through systematic programs of research. *Management Inform. Systems Quart.* 41(2): 335–351.

Omohundro S (2014) Autonomous technology and the greater human good. *J. Experiment. Theoretical Artificial Intelligence* 26(3):303–315.

Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Preprint, submitted September 29, https://arxiv.org/abs/cs/0409058.

Pearl J (2019) The seven tools of causal inference, with reflections on machine learning. *Comm. ACM* 62(3):54–60.

R Core Team (2021) R: A language and environment for statistical computing. https://www.R-project.org/.

Raisch S, Krakowski S (2020) Artificial intelligence and management: The automation–augmentation paradox. *Acad. Management Rev.* 46(1):192–210.

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. Krishnapuram B, Shah M, eds. *Proc. 22nd ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 1135–1144.

Riessman CK (2011) What's different about narrative inquiry? Cases, categories and contexts. Silverman D, ed. *Qualitative Research: Issues of Theory, Method, and Practice*, 3rd ed. (Sage, London), 310–330.

Roth WM (2005) Making classifications (at) work: Ordering practices in science. *Soc. Stud. Sci.* 35(4):581–621.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Natural Machine Intelligence* 1(5):206–215.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, et al. (2015) Imagenet large scale visual recognition challenge. *Internat. J. Comput. Vision* 115:211–252.

Sapir A, Drori I, Ellis S (2016) The practices of knowledge creation: Collaboration between peripheral and core occupational communities. *Eur. Managment Rev.* 13(1):19–36.

Schreier M (2012) *Qualitative Content Analysis in Practice* (Sage Publications, London).

Schwartz DG (1995) *Cooperating Heterogeneous Systems* (Kluwer Academic Publishers, Alphen aan den Rijn, Netherlands).

Schwartz DG, Te'eni D (2000) Tying knowledge to action with kMail. *IEEE Intelligent Systems Their Appl.* 15(3):33–39.

Schwartz DG, Yahav I (2021) Knowledge contribution diagrams for design science research: A novel graphical technique. *Proc. Internat. Conf. on Design Sci. Res. in Inform. Systems and Tech.*, Part of the Lecture Notes in Computer Science (Springer, Berlin), 174–187.

Seidel S, Berente N, Lindberg A, Lyytinen K, Nickerson JV (2018) Autonomous tools and design: A triple-loop approach to human-machine learning. *Comm. ACM* 62(1):50–57.

Sein MK, Henfridsson O, Purao S, Rossi M, Lindgren R (2011) Action design research. *Management Inform. Systems Quart.* 35(1):37–56.

Sengupta K, Te'eni D (1993) Cognitive feedback in GDSS: Improving control and convergence. *Management Inform. Systems Quart.* 17(1):87–113.

Shrestha YR, Ben-Menahem SM, Von Krogh G (2019) Organizational decision-making structures in the age of artificial intelligence. *California Management Rev.* 61(4):66–83.

Silverman G, Sommer U (2019) Prevalent sentiments of the concept of jihad in the public commentsphere. *Stud. Conflict Terrorism* 45(7):579–607.

So C (2020a) Human-in-the-loop design cycles: A process framework that integrates design sprints, agile processes, and machine learning with humans. *Proc. 1st Internat. Conf. on Artificial Intelligence in HCI, AI-HCI*, Part of the Lecture Notes in Computer Science (Springer, Berlin), 136–145.

So C (2020b) Understanding the prediction mechanism of sentiments by XAI visualization. *Proc. 4th Internat. Conf. on Natural Language Processing and Inform. Retrieval* (ACM, New York), 75–80.

Son JY, Goldstone RL (2009) Contextualization in perspective. *Cognitive Instruction* 27(1):51–89.

Sturm T, Gerlach JP, Pumplun L, Mesbah N, Peters F, Tauchert C, Nan N, et al. (2021) Coordinating human and machine learning for effective organizational learning. *Management Inform. Systems Quart.* 45(3):1581–1602.

Suchman LA (1987) *Plans and Situated Actions: The Problem of Human-Machine Communication* (Cambridge University Press, Cambridge, UK).

Suchman LA (2007) *Human-Machine Reconfigurations: Plans and Situated Actions* (Cambridge University Press, Cambridge, UK).

Tausczik YR, Pennebaker JW (2009) The psychological meaning of words: LIWC and computerized text analysis methods. *J. Language Soc. Psych.* 29(1):24–54.

Te'eni D (2001) Review: A cognitive-affective model of organizational communication for designing IT. *Management Inform. Systems Quart.* 25(2):251–312.

Van den Broek E, Sergeeva A, Huysman M (2021) When the machine meets the expert: An ethnography of developing AI for hiring. *Management Inform. Systems Quart.* 45(3):1557–1580.

Vanides J, Yin Y, Tomita M, Ruiz-Primo MA (2005) Concept maps. *Sci. Scope* 28(8):27–31.

Vimalkumar M, Gupta A, Sharma D, Dwivedi Y (2021) Understanding the effect that task complexity has on automation potential and opacity: Implications for algorithmic fairness. *AIS Trans. Human-Comput. Interactions* 13(1):104–129.

Vygotsky LS, Rice E (1978) *Mind in Society: The Development of Higher Mental Processes* (Harvard University Press, Cambridge, MA).

Wang X, Kapanipathi P, Musa R, Yu M, Talamadupula K, Abdelaziz I, Chang M, et al. (2019) Improving natural language inference using external knowledge in the science questions domain. *Proc. Conf. AAAI Artificial Intelligence* 33(1):7208–7215.

Weick KE, Sutcliffe KM, Obstfeld D (2005) Organizing and the process of sensemaking. *Organ. Sci.* 16(4):409–421.

Woods DD, Hollnagel E (2006) *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering* (CRC Press, Boca Raton, FL).

Zagalsky A, Te'eni D, Yahav I, Schwartz DG, Silverman G, Cohen D, Mann Y, Lewinsky D (2021) The design of reciprocal learning between human and artificial intelligence. Nichos J, ed. *Proc. ACM Human-Comput. Interaction*, 5(CSCW2) (Association for Computing Machinery, New York), 1–36.

Zamani ED, Griva A, Spanaki K, O'Raghallaigh P, Sammon D (2021) Making sense of business analytics in project selection and prioritisation: Insights from the start-up trenches. *Inform. Tech. People.*

Zerilli J, Knott A, Maclaurin J, Gavaghan C (2019) Algorithmic decision-making and the control problem. *Minds Machine* 29(4): 555–578.

Zhao R, Mao K (2018) Fuzzy bag-of-words model for document representation. *IEEE Trans. Fuzzy Systems* 26(2):794–804.