Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?

Sriraam Natarajan[1], **Saurabh Mathur**[1*], **Sahil Sidheekh**[1*], **Wolfgang Stammer**[2,3], **Kristian Kersting**[2,3]

[1] University of Texas at Dallas, [2] Technical University of Darmstadt,
[3] Hessian Center for Artificial Intelligence (hessian.ai), Darmstadt, Germany

Abstract

Human-in-the-loop (HIL) systems have emerged as a promising approach for combining the strengths of data-driven machine learning models with the contextual understanding of human experts. However, a deeper look into several of these systems reveals that calling them HIL would be a misnomer, as they are quite the opposite, namely AI-in-the-loop (AI^2L) systems: the human is in control of the system, while the AI is there to support the human. We argue that existing evaluation methods often overemphasize the machine (learning) component's performance, neglecting the human expert's critical role. Consequently, we propose an AI^2L perspective, which recognizes that the human expert is an active participant in the system, significantly influencing its overall performance. By adopting an AI^2L approach, we can develop more comprehensive systems that faithfully model the intricate interplay between the human and machine components, leading to more effective and robust AI systems.

Introduction

Since the time of the "advice taker" (McCarthy 1959, 1968), there has been a significant interest in building human-allied AI systems. Different paradigms and different techniques such as active learning (Settles 2009), knowledge-based learning (Towell and Shavlik 1994; Fung, Mangasarian, and Shavlik 2002), explanatory interactive learning (Schramowski et al. 2020; Stammer, Schramowski, and Kersting 2021), advice-taking (DeJong and Mooney 1986; Baffes and Mooney 1996; Odom et al. 2015), weak/distant supervision (Natarajan et al. 2014; Ratner et al. 2017), human feedback (Maclin et al. 2005; Wiewiora, Cottrell, and Elkan 2003; Ng, Harada, and Russell 1999; Griffith et al. 2013) and preference elicitation (Boutilier 2002; Boutilier et al. 2004; Chen and Pu 2012; Toni et al. 2024) etc., to name a few, have been developed for this important and challenging task. Many of these directions have been presented under the umbrella of *Human-in-the-loop* (HIL) systems. We take a deeper look at these systems and ask the question if they are truly HIL systems.

To understand the difference, let us consider two simple examples:

1. An AI system that recommends content to users (say videos). If a human intervenes in such a system, they provide feedback/guidance by either correcting inappropriate content (as a trusted ally) or by providing malicious/inappropriate content (as an adversary). In either case, the AI agent optimizes its internal function, considers the human feedback, and decides on the appropriate action (in this case, showing the relevant content).
2. As a second example, consider an AI system that assists a physician who is treating a diabetic patient for a knee injury and prescribes oral steroids to mitigate pain. AI could now intervene based on its internal objective function, domain constraints, and knowledge and suggest that since the patient has diabetes, the physician should reconsider their recommendation. The physician can then inform the system that the patient is in acute pain and reducing that is important or, in contrast, that the patient did not inform the physician and hence will change her prescription.

Indeed, in either of these cases, the AI system interacts with the human expert, assimilates knowledge, updates its constraints, makes internal computations, and then provides suggestions. However, although these two systems appear quite similar at the outset, there is a crucial difference in the role of decision-making authority and control. In the former case, AI is in charge of decision-making and takes additional inputs from the human expert. Arguably, these can be "richer" inputs than treating the human as a "mere labeler". Still, the human is not the decision-maker, while the AI actually is. In the latter case, the human is in control of the full system. The presence of AI inside this system only makes the process more efficient and possibly more effective. However, the full system exists independent of the presence of AI. This difference is critical.

Consider the problem of evaluation. Clearly, evaluating a system based only on its performance (say accuracy or some other function of precision/recall) will benefit the HIL system but not the AI^2L system, as argued by van Amsterdam et al. (2024). More importantly, during deployment, the issues that the AI^2L systems must address and reason with can be significantly different from HIL systems.

In the rest of this blue sky paper, we first present these two systems in greater detail showing their similarities and differences. We argue strongly that when designing systems

[*] These authors contributed equally.

that operate in the presence of human experts, the designers of these systems must clearly understand the conditions under which these systems operate and then decide whether a HIL or AI^2L system is appropriate for the task at hand. After all, there is no one ring to rule them all!

Human-in-the-loop

The Human-in-the-loop paradigm for developing AI systems typically treats humans either as data-labeling oracles (Settles 2009) or as a source of domain knowledge (Mosqueira-Rey et al. 2023). Humans primarily function as oracles in the Active Learning (AL, see Settles (2009)) paradigm or as weak supervisors (Ratner et al. 2017), providing labels for unlabeled data instances the model finds uncertain, ambiguous, or missing. These approaches are valuable for domains with large amounts of unlabeled data and for which annotation is costly or time-consuming. While the system controls the learning process by selecting which instances are presented to the human for labeling, AL aims to improve model accuracy with fewer training examples. However, this approach relies on the assumption that humans prefer acting as efficient labeling machines, falling short of making *human use of human beings* (Wiener 1988).

Machine teaching (MT, see Simard et al. (2017)) focuses on making the teachers who build machine learning models more effective, rather than improving just the learning algorithms. Over the past two decades, however, most research has centered on developing powerful (deep) learning algorithms for handling abundant data (LeCun, Bengio, and Hinton 2015). However, as machine learning expands to address more varied and often rather short-term tasks, the scarcity and cost of skilled teachers have become limiting factors. Inspired by the evolution of programming in the 1980s and 1990s, machine teaching emphasizes principles like problem decomposition, modularity, and process discipline. It draws parallels with programming, highlighting version control, semantic data exploration, and the expressiveness of teaching languages as key aspects of effective machine teaching.

Specifically, the MT paradigm posits humans as teachers who guide machine learning models to acquire specific knowledge. This allows domain experts to create effective models in the absence of large data sets without deep ML expertise. Such knowledge-intensive learning methods have a long history in AI, from John McCarthy's work in the 1960s (McCarthy 1968) to explanation-based learning (DeJong and Mooney 1986), theory refinement (Baffes and Mooney 1996), and inductive logic programming and relational learning (Muggleton and Raedt 1994; Raedt 2008). The key motivation is that experts have extensive knowledge in their respective fields, which many data-driven ML techniques (Mitchell 1997; Steinwart and Christmann 2008; Natarajan et al. 2015) do not fully exploit.

In fact, the use of advice in various forms has produced successful algorithms, particularly in reinforcement learning (Maclin et al. 2005; Wiewiora, Cottrell, and Elkan 2003; Ng, Harada, and Russell 1999), where advice is used as reward shaping. In supervised learning, advice is typically provided as feature selection or inductive bias on initial models. In graphical models, advice is used as an initial structure that is refined (Heckerman 1998). On the other hand, knowledge-based neural networks (Towell and Shavlik 1994) and support vector machines (Fung, Mangasarian, and Shavlik 2002; Kunapuli et al. 2010), inverse RL (Kunapuli et al. 2013), relational models (Odom et al. 2015; Odom and Natarajan 2018, 2016) and probabilistic model learning (Altendorf, Restificar, and Dietterich 2005; de Campos, Tong, and Ji 2008; Yang and Natarajan 2013; Kokel et al. 2020; Mathur, Gogate, and Natarajan 2023; Mathur, Antonucci, and Natarajan 2024) have explored combining knowledge and data to handle systematic noise. While adaptation specifics may differ, all of these methods can take advice as Horn clauses, convert them to their corresponding representation, and learn by using them as constraints.

The common denominator in all of these systems is that the AI module is in control of the decision process and the human inputs are essentially used to "guide" the model to a better (possibly local) optimum. This scenario is explored in the LHS of Figure 1. The performance of HIL systems is typically measured from the system's perspective and is based on accuracy, precision, recall, or a function of these metrics. Issues of trust (due to inherent biases), and credibility are important challenges in these systems. These mainly stem from the human expert's biases and data biases and include, but are not limited to confirmation bias, conformity bias, attribution bias, affinity bias, halo effect, cognitive bias, and racial and gender bias to name a few. A common issue of HIL systems is thus the danger of manipulation by an adversary that provides incorrect advice, ultimately requiring the system to model the credibility of the human experts to make effective decisions.

AI-in-the-loop

Many applications of AI and machine learning involve interactions with humans. Humans may provide input to a learning algorithm, including input in the form of labels, demonstrations, corrections, rankings, or evaluations. They could give such input while observing the algorithm's outputs, potentially in the form of feedback, predictions, or demonstrations. However, although humans are arguably an integral part of the learning process, traditional machine learning systems are agnostic to the fact that inputs/outputs are from/-for humans. In fact, machine learning is often conceived — in particular in applications of other scientific disciplines such as medicine — in a very impersonal way, with algorithms working autonomously on passively collected data.

In contrast, interactive machine learning (IML, Fails and Olsen Jr (2003); Amershi et al. (2014); Michael, Acklin, and Scheuerman (2020); Ware et al. (2001); Wang (2019); Teso et al. (2023)) represents a shift toward greater human involvement and shared control in the learning process. Humans can assume various roles in IML, from domain experts and data scientists to non-expert users. This flexibility allows for a more dynamic interplay between humans and machines, assigning tasks based on individual strengths and capabilities. Unlike the algorithm-centric focus of AL,

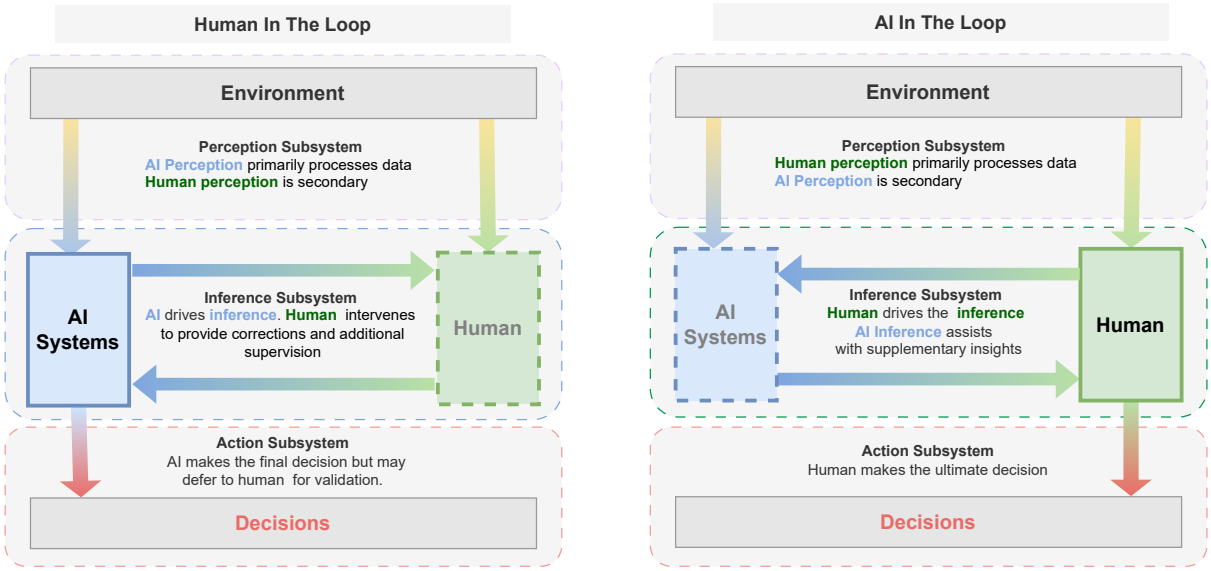

Figure 1: A comparison of human in the loop (left) and AI in the loop (right) systems. In human-in-the-loop systems, AI systems drive the inference and decision-making process, but humans intervene to provide corrections and supervision. In AI-in-the-loop systems, humans make the ultimate decisions, while AI systems assist with perception, inference, and action.

IML systems require a human-centered approach to evaluation that incorporates additional judgments such as calibration, fairness, and explainability alongside traditional performance metrics that merely measure the system's conformity to past data. Beyond their direct involvement in training, humans are also the ultimate users of AI systems. This requires evaluating AI systems not only for their functionality but also for their usability and usefulness to human users.

In keeping with the paradigm shift of IML systems, we argue for a change in thinking about these systems. Specifically, consider the RHS of Figure 1. It can be easily seen that while the system is very similar to the HIL system, there are crucial differences. While in HIL systems AI is in control, in AI^2L, the human is at the center of the system and fully in control. Despite these critical differences, *e.g.*, in the deployment of safe AI-in-the-loop systems, we observe that many of the existing literature simply consider these two systems to belong to one group. However, we strongly argue for their separation and will highlight their differences next.

First, consider the issues of reliability and trust in these systems. The biases in AI^2L systems are mainly due to algorithmic and model biases that are reflective of the data bias. Moreover, trust issues in these systems are drastically different from those of HIL systems. While the credibility assessment of the human teacher is crucial in HIL, transparency of the system, its explainability, and interpretability are crucial in AI^2L systems (Ross, Hughes, and Doshi-Velez 2017; Teso and Kersting 2019; Lipton 2018; Rudin 2019; Schramowski et al. 2020; Stammer et al. 2024). Moreover, trust is much more nuanced in AI^2L systems than HIL systems because the human user is in control and is unlikely to trust a system if it does not align with their expectations. Hence, the user's confirmation bias could potentially be reflected in their willingness to trust the underlying AI^2L system. From the perspective of credibility in the context of AI^2L, instead of assigning credibility to humans (as in HIL systems), typically AI^2L systems compute the credibility of data sources, for example, the different modalities or knowledge bases from which the data are being extracted.

Second, the evaluations of these systems are human-centric and are mostly aligned with the broader goals of the environment in which they operate. Although metrics such as Precision, Recall, Accuracy, and F-scores etc. are still relevant in understanding the performance of the AI system, arguably much more emphasis should be placed on the impact on the human who is at the center of the system. Hence, ablation studies are more important in such systems to evaluate the impact of the different components on the overall system, for instance, on specific health outcomes. While in HIL systems, ablation studies are useful, in AI^2L systems, they are essential. Typically, some other important aspects of the evaluation of these AI^2Lsystems are the interpretability, explainability(Sreedharan, Kulkarni, and Kambhampati 2022), interactive capabilities (Zahedi et al. 2023), and generalizability of these systems (Wüst et al. 2024).

Above all, the most important consideration concerns the system itself. Is the AI system necessary in this task that is already performed by the human? If so, what is the potential impact of the AI system, efficiency or efficacy, or both? What are the potential hazards of using an AI system in this task? How can the improvements in the systems be measured objectively? Furthermore, are the biases due to the model or the data? How credible are the data sources that helped create the AI^2L system? These questions must be answered

Table 1: Examples of subtasks across various domains and their classification as **Automate (HIL)** or **Collaborate** (AI^2L).

Domain	**Task**	**Automate (HIL) or Collaborate (AI^2L)?**	**Description**
Medicine	Early diagnosis of Alzheimer's disease	Automate	AI analyzes patient data (e.g., MRI) and detects anomalies.
	Treatment plan formulation	Collaborate	Physician selects and tailors the final treatment plan from a set of candidates generated by AI.
Automobile	Route planning from source to destination	Automate	AI computes optimal routes using real-time traffic and weather data, requiring minimal human input.
	Driving in high-density, urban environment	Collaborate	Human drivers navigate complex traffic, assisted by AI for tasks like collision avoidance, lane change and adaptive cruise control.
Logistics	Shipping cost forecasting based on historical data	Automate	AI uses prior data to predict shipping costs; human manager can correct previous mistakes, provide additional context and advice.
	Inventory management	Collaborate	Human decides on procurement strategy based on AI's estimates of stock/reorder points.
Manufacturing	Detecting known quality issues in product	Automate	AI automatically detects product defects; human inspectors confirm corrections if needed.
	Quality assurance and compliance	Collaborate	Humans design quality assurance strategies based on quality issue patterns identified by AI.
Finance	Fraud detection	Automate	AI analyzes the data and flags suspicious activities/transactions; humans confirm.
	Investment advisory	Collaborate	Human advisor decides on the final strategy using AI-provided market analyses and recommendations.
Education	Automated grading of assignments	Automate	AI gives feedback on assignments based on commonly seen mistakes (e.g., software bugs); human instructor provides advice, sets grading schema and policies, and reviews uncertain cases.
	Curriculum Updates	Collaborate	Teachers update and adapt lessons to student needs based on AI's analyses of student learning trajectories across time.

deliberately before the system can be deployed.

This is the crux of our argument – *instead of considering every system in which a human is present to be a HIL system, it is imperative to understand the type of system, their evaluation criteria, and potential implications of its deployment in greater detail.*

In fact, AI^2L systems are related to the vision of bridging explainable and advisable AI and achieving a human-AI symbiosis (Zahedi and Kambhampati 2021; Sreedharan, Kulkarni, and Kambhampati 2022; Kambhampati et al. 2022; Zahedi et al. 2023; Kambhampati 2020). Both emphasize that the human and the computer are both in the loop, and AI becomes a co-adaptive process, in which a human is changing AI behavior, but the human also adapts to use AI more effectively and adapts their data and goals in response to what is learned using machine learning. AI^2L systems, however, emphasize the need to move beyond the train/test evaluation paradigm of static, non-contextualised benchmarks, toward user- or even population-specific metrics and evaluation protocols close to the real-life requirements of society.

Discussion

HIL and AI2L systems differ in three key aspects, namely, control, source of bias, and evaluation. The first difference is that HIL systems are generally autonomous AI agents that might seek specific help from humans, while AI2L systems constitute an intervention in a human decision-making process. The AI component in AI2L presents the human with a summary of information synthesized from multiple sources, a set of possible allowable actions, and their possible consequences (e.g., a human selecting a single decision from a set of MPE solutions). This difference in control configurations results in differing sources of bias. While HIL systems are primarily vulnerable to bias in historical data and domain knowledge used in model construction, AI2L systems are also vulnerable to biases arising from human interpretation of the AI's output. Finally, while HIL systems are evaluated using AI-centered metrics such as accuracy, precision, and recall, AI2L systems require a more holistic approach to evaluation, taking into account the human-AI interaction, the overall goals of the decision-making process, and considerations such as fairness that cannot be fully quantified. Table 1 presents a few concrete examples from

diverse domains, grounding the distinctions between HIL and AI^2L systems in practical, real-world contexts. It categorizes sample subtasks across areas such as Medicine, Automobile, Logistics, Manufacturing, Finance, and Education, illustrating whether they are best addressed through automation with HIL oversight or through collaborative engagement via AI^2L systems.

The choice of HIL or AI^2L perspective influences the aspects of a system that are abstracted away during design and evaluation. Designing systems as HIL when they should be understood as AI^2L can result in abstraction errors (Selbst et al. 2019), allow for modeling the conscious and unconscious biases that arise due to humans, evaluate the system on incorrect or inappropriate criteria, or have serious consequences after deployment. For instance, using an HIL system to regulate an exceedingly complex stochastic system like the human body (Beer 1967) might overlook crucial contextual details necessary for clinical decision-making. In contrast, the AI^2L perspective recasts system deployment as an intervention in existing processes, allowing evaluation strategies to be more closely aligned with end goals such as improving health outcomes (van Amsterdam et al. 2024) and minimizing harmful social outcomes (Mohla, Bagh, and Guha 2021).

In summary, the appropriate problem domains for the HIL and AI^2L systems are typically not separated but nested. Automation is most effective in well-defined contexts, while human intervention is most needed in not yet defined or undefinable contexts. Hence, zooming in on a domain would result in a HIL problem and zooming out would give us an AI^2L problem, *e.g.,* identifying drug-drug interactions is a reasonably well understood context, making it appropriate for HIL while general medical diagnosis is not as clearly defined, making it more appropriate for AI^2L. Additionally, while effective software engineering requires active human decision-making (Johnson and Menzies 2024), automating some well-defined sub-problems such as static analysis and vulnerability detection (Yadavally et al. 2023) can help reduce the software engineer's cognitive load and lead to better quality software.

While our discussions are motivated from the perspective of supervised learning, the frameworks are agnostic to the type of learning performed. Our arguments for the difference in the two systems apply directly to unsupervised learning, reinforcement learning, planning, continual learning, and meta-learning to name a few. While specific adaptations differ, the idea of humans in the center or AI in the center applies broadly across these different settings, see e.g. (Delfosse et al. 2024). Or consider current foundation models trained in a self-supervised fashion. While AI^2L systems focus on AI-supported human agency, current foundation models lend themselves more easily to HIL settings where AI is the primary actor, and humans intervene to monitor or enhance results. Their ability to generalize across tasks and perform with minimal additional training allows humans to take on roles of oversight and feedback rather than constant, direct involvement—the user still gives a thumbs up or down on the text generated by a large language model. Although they may pick up information about how to collaborate with humans, without an understanding of the user's goals, beliefs, or uncertainties, even foundation models are likely to remain reactive rather than truly collaborative partners. By improving their "theory of mind", however, foundation models could offer more contextual, meaningful suggestions that integrate with human workflows and align better with social values and the evolving needs of human users.

In short, moving from HIL to AI^2L is likely to help build AI systems where AI truly enhances human expertise, resulting in smarter, more resilient solutions that thrive on collaboration, not automation. Doing so, however, requires the AI community to rethink its evaluation methodology.

Acknowledgements

SN sincerely thanks Rao Kambhampati for his insightful discussions that led to a deeper dive into the differences between HIL and AI^2L systems. SN, SM, and SS gratefully acknowledge the generous support by AFOSR award FA9550-23-1-0239, the ARO award W911NF2010224 and the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) award HR001122S0039. KK and WS acknowledge that they benefited from the "ML2MT" project (Volkswagen Stiftung) and the 3AI project from the Hessian Ministry of Science and Arts (HMWK). KK and WS also benefited from the European Project "Tango" (Grant Agreement no. 101120763); the views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA); neither the European Union nor the granting authority can be held responsible for them.

References

Altendorf, E.; Restificar, A. C.; and Dietterich, T. G. 2005. Learning from Sparse Data by Exploiting Monotonicity Constraints. In *UAI*, 18–26. AUAI Press.

Amershi, S.; Cakmak, M.; Knox, W. B.; and Kulesza, T. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4): 105–120.

Baffes, P. T.; and Mooney, R. J. 1996. A novel application of theory refinement to student modeling. In *Proceedings of the thirteenth national conference on Artificial intelligence-Volume 1*, 403–408.

Beer, S. 1967. *Cybernetics and management*. London: English Universities Press.

Boutilier, C. 2002. A POMDP Formulation of Preference Elicitation Problems. In *AAAI/IAAI*, 239–246. AAAI Press / The MIT Press.

Boutilier, C.; Brafman, R. I.; Domshlak, C.; Hoos, H. H.; and Poole, D. 2004. CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements. *J. Artif. Intell. Res.*, 21: 135–191.

Chen, L.; and Pu, P. 2012. Critiquing-based recommenders: survey and emerging trends. *User Model. User Adapt. Interact.*, 22(1-2): 125–150.

de Campos, C. P.; Tong, Y.; and Ji, Q. 2008. Constrained Maximum Likelihood Learning of Bayesian Networks for Facial Action Recognition. In *ECCV (3)*, volume 5304 of *Lecture Notes in Computer Science*, 168–181. Springer.

DeJong, G.; and Mooney, R. J. 1986. Explanation-Based Learning: An Alternative View. *Mach. Learn.*, 1(2): 145–176.

Delfosse, Q.; Sztwiertnia, S.; Rothermel, M.; Stammer, W.; and Kersting, K. 2024. Interpretable Concept Bottlenecks to Align Reinforcement Learning Agents. *NeurIPS*.

Fails, J. A.; and Olsen Jr, D. R. 2003. Interactive Machine Learning. In *International Conference on Intelligent User Interfaces*, 39–45.

Fung, G.; Mangasarian, O. L.; and Shavlik, J. W. 2002. Knowledge-Based Support Vector Machine Classifiers. In *NIPS*, 521–528. MIT Press.

Griffith, S.; Subramanian, K.; Scholz, J.; Jr., C. L. I.; and Thomaz, A. L. 2013. Policy Shaping: Integrating Human Feedback with Reinforcement Learning. In *NIPS*, 2625–2633.

Heckerman, D. 1998. A tutorial on learning with Bayesian networks. *Learning in graphical models*, 301–354.

Johnson, B.; and Menzies, T. 2024. AI Over-Hype: A Dangerous Threat (and How to Fix It). *IEEE Software*, 41(6): 131–138.

Kambhampati, S. 2020. Challenges of Human-Aware AI Systems AAAI Presidential Address. *AI Mag.*, 41(3): 3–17.

Kambhampati, S.; Sreedharan, S.; Verma, M.; Zha, Y.; and Guan, L. 2022. Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems. In *AAAI*, 12262–12267. AAAI Press.

Kokel, H.; Odom, P.; Yang, S.; and Natarajan, S. 2020. A Unified Framework for Knowledge Intensive Gradient Boosting: Leveraging Human Experts for Noisy Sparse Domains. In *AAAI*, 4460–4468. AAAI Press.

Kunapuli, G.; Bennett, K. P.; Shabbeer, A.; Maclin, R.; and Shavlik, J. W. 2010. Online Knowledge-Based Support Vector Machines. In *ECML/PKDD (2)*, volume 6322 of *Lecture Notes in Computer Science*, 145–161. Springer.

Kunapuli, G.; Odom, P.; Shavlik, J. W.; and Natarajan, S. 2013. Guiding Autonomous Agents to Better Behaviors through Human Advice. In *ICDM*, 409–418. IEEE Computer Society.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature*, 521(7553): 436–444.

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

Maclin, R.; Shavlik, J. W.; Torrey, L.; Walker, T.; and Wild, E. W. 2005. Giving Advice about Preferred Actions to Reinforcement Learners Via Knowledge-Based Kernel Regression. In *AAAI*, 819–824. AAAI Press / The MIT Press.

Mathur, S.; Antonucci, A.; and Natarajan, S. 2024. Knowledge Intensive Learning of Credal Networks. In *UAI*. PMLR.

Mathur, S.; Gogate, V.; and Natarajan, S. 2023. Knowledge Intensive Learning of Cutset Networks. In *UAI*, volume 216 of *Proceedings of Machine Learning Research*, 1380–1389. PMLR.

McCarthy, J. 1959. Programs with common sense. In *Proc. Teddington Conference on the Mechanization of Thought Processes, 1959*, 75–91.

McCarthy, J. 1968. Programs with Common Sense. In *Semantic information processing*. The MIT Press.

Michael, C. J.; Acklin, D.; and Scheuerman, J. 2020. On interactive machine learning and the potential of cognitive feedback. *arXiv preprint arXiv:2003.10365*.

Mitchell, T. 1997. *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc., 1 edition. ISBN 0070428077, 9780070428072.

Mohla, S.; Bagh, B.; and Guha, A. 2021. A material lens to investigate the gendered impact of the ai industry. In *IJCAI 2021 Workshop on AI for Social Good*.

Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; and Fernández-Leal, Á. 2023. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, 56(4): 3005–3054.

Muggleton, S. H.; and Raedt, L. D. 1994. Inductive Logic Programming: Theory and Methods. *J. Log. Program.*, 19/20: 629–679.

Natarajan, S.; Kersting, K.; Khot, T.; and Shavlik, J. 2015. *Boosted statistical relational learners: From benchmarks to data-driven medicine*. Springer.

Natarajan, S.; Picado, J.; Khot, T.; Kersting, K.; Ré, C.; and Shavlik, J. W. 2014. Effectively Creating Weakly Labeled Training Examples via Approximate Domain Knowledge. In *ILP*, volume 9046 of *Lecture Notes in Computer Science*, 92–107. Springer.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, 278–287.

Odom, P.; Khot, T.; Porter, R. B.; and Natarajan, S. 2015. Knowledge-Based Probabilistic Logic Learning. In *AAAI*, 3564–3570. AAAI Press.

Odom, P.; and Natarajan, S. 2016. Active Advice Seeking for Inverse Reinforcement Learning. In *AAMAS*, 512–520. ACM.

Odom, P.; and Natarajan, S. 2018. Human-Guided Learning for Probabilistic Logic Models. *Frontiers Robotics AI*, 5: 56.

Raedt, L. D. 2008. *Logical and relational learning*. Cognitive Technologies. Springer. ISBN 978-3-540-20040-6.

Ratner, A.; Bach, S. H.; Ehrenberg, H. R.; Fries, J. A.; Wu, S.; and Ré, C. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proc. VLDB Endow.*, 11(3): 269–282.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In Sierra, C., ed., *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 2662–2670.

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

Schramowski, P.; Stammer, W.; Teso, S.; Brugger, A.; Herbert, F.; Shao, X.; Luigs, H.; Mahlein, A.; and Kersting, K. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2(8): 476–486.

Selbst, A. D.; danah boyd; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *FAT*, 59–68. ACM.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Simard, P. Y.; Amershi, S.; Chickering, D. M.; Pelton, A. E.; Ghorashi, S.; Meek, C.; Ramos, G.; Suh, J.; Verwey, J.; Wang, M.; et al. 2017. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742*.

Sreedharan, S.; Kulkarni, A.; and Kambhampati, S. 2022. *Explainable human-AI interaction: A planning perspective*. Springer Nature.

Stammer, W.; Schramowski, P.; and Kersting, K. 2021. Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting With Their Explanations. In *CVPR*, 3619–3629. Computer Vision Foundation / IEEE.

Stammer, W.; Wüst, A.; Steinmann, D.; and Kersting, K. 2024. Neural Concept Binder. *NeurIPS*.

Steinwart, I.; and Christmann, A. 2008. Support Vector Machines, Book. *Information Science and Statistics*.

Teso, S.; Alkan, Ö.; Stammer, W.; and Daly, E. 2023. Leveraging explanations in interactive machine learning: An overview. *Frontiers Artif. Intell.*, 6.

Teso, S.; and Kersting, K. 2019. Explanatory Interactive Machine Learning. In Conitzer, V.; Hadfield, G. K.; and Vallor, S., eds., *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society(AIES)*, 239–245.

Toni, G. D.; Viappiani, P.; Teso, S.; Lepri, B.; and Passerini, A. 2024. Personalized Algorithmic Recourse with Preference Elicitation. *Trans. Mach. Learn. Res.*

Towell, G. G.; and Shavlik, J. W. 1994. Knowledge-based artificial neural networks. *Artificial intelligence*, 70(1-2): 119–165.

van Amsterdam, W. A. C.; de Jong, P. A.; Verhoeff, J. J. C.; Leiner, T.; and Ranganath, R. 2024. From algorithms to action: improving patient care requires causality. *BMC Medical Informatics Decis. Mak.*, 24(3): 111.

Wang, G. 2019.

Ware, M.; Frank, E.; Holmes, G.; Hall, M. A.; and Witten, I. H. 2001. Interactive machine learning: letting users build classifiers. *Int. J. Hum. Comput. Stud.*, 55(3): 281–292.

Wiener, N. 1988. *The human use of human beings: Cybernetics and society*. 320. Da capo press.

Wiewiora, E.; Cottrell, G. W.; and Elkan, C. 2003. Principled Methods for Advising Reinforcement Learning Agents. In *ICML*, 792–799. AAAI Press.

Wüst, A.; Stammer, W.; Delfosse, Q.; Dhami, D. S.; and Kersting, K. 2024. Pix2Code: Learning to Compose Neural Visual Concepts as Programs. *Uncertainty in Artificial Intelligence (UAI)*.

Yadavally, A.; Nguyen, T. N.; Wang, W.; and Wang, S. 2023. (Partial) Program Dependence Learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2501–2513. IEEE.

Yang, S.; and Natarajan, S. 2013. Knowledge Intensive Learning: Combining Qualitative Constraints with Causal Independence for Parameter Learning in Probabilistic Models. In *ECML/PKDD (2)*, volume 8189 of *Lecture Notes in Computer Science*, 580–595. Springer.

Zahedi, Z.; and Kambhampati, S. 2021. Human-AI symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990*.

Zahedi, Z.; Verma, M.; Sreedharan, S.; and Kambhampati, S. 2023. Trust-Aware Planning: Modeling Trust Evolution in Iterated Human-Robot Interaction. In *HRI*, 281–289. ACM.