Learning Paradigms for Hybrid Decision-Making Systems

CLARA PUNZI, Faculty of Sciences, Scuola Normale Superiore, Pisa, Italy and Department of Computer Science, University of Pisa, Pisa, Italy
ROBERTO PELLUNGRINI, Faculty of Sciences, Scuola Normale Superiore, Pisa, Italy
MATTIA SETZU, Department of Computer Science, University of Pisa, Pisa, Italy
FOSCA GIANNOTTI, Scuola Normale Superiore, Pisa, Italy
DINO PEDRESCHI, Department of Computer Science, University of Pisa, Pisa, Italy

The rapid integration of AI systems into high-stakes domains has revealed persistent issues of user distrust, algorithmic aversion, and over-reliance, highlighting the need for decision-making frameworks in which humans and machines synergistically collaborate towards the solution of the task. Hybrid Decision-Making Systems (HDMS) have emerged as a paradigm where humans and AI jointly contribute to the same task, leveraging and integrating human strengths like domain expertise, contextual understanding and flexible reasoning, alongside machines' computational power. This survey offers a structured overview of learning paradigms for HDMS, with a particular focus on uncertainty-driven abstention mechanisms, which determine when an AI system should act autonomously or when it should call for human intervention. We formalise and compare algorithmic approaches that embed machine learning models with the capacity to "know what they don't know", analysing how abstention policies and system architectures integrate human expertise into the decision pipeline. Beyond abstention, we examine frameworks that support direct human–machine interaction during and after the learning process, outlining emerging approaches that foster bidirectional collaboration between humans and AI. Building on this analysis, we propose a taxonomy of three learning paradigms characterising progressively tighter human–machine integration.

CCS Concepts: • **Computing methodologies** → **Learning paradigms**; • **Human-centered computing** → *HCI theory, concepts and models.*

Additional Key Words and Phrases: Hybrid decision making, Learning to Abstain, Human-AI collaboration

1 Introduction

The rapid adoption of AI systems (including LLMs) across high-stakes fields like medicine and finance has highlighted issues of user distrust [41], misuse, disuse and amplified societal biases by human actors [91], with the risk of incurring undesired patterns of user behaviour, such as algorithmic aversion [41] and over-reliance [101]. In response, the "Human-in-the-Loop" (HITL) paradigm has emerged, where AI acts as a collaborator to human experts, who retain final decision responsibility. This synergistic framework, which acknowledges the shared nature of cognition, enhances the quality, transparency, and fairness of decisions of the decision-making process [22, 113]. We define these as Hybrid Decision-Making Systems (HDMS), where human and machine agents jointly contribute to the same task. HDMS leverages complementary strengths: humans provide domain expertise, contextual understanding, commonsense reasoning [26] and lateral thinking [30], while machines

Authors' Contact Information: Clara Punzi, Faculty of Sciences, Scuola Normale Superiore, Pisa, Toscana, Italy and Department of Computer Science, University of Pisa, Pisa, Italy; e-mail: clara.punzi@sns.it; Roberto Pellungrini, Faculty of Sciences, Scuola Normale Superiore, Pisa, Italy; e-mail: roberto.pellungrini@sns.it; Mattia Setzu, Department of Computer Science, University of Pisa, Pisa, Italy; e-mail: mattia.setzu@unipi.it; Fosca Giannotti, Scuola Normale Superiore, Pisa, Toscana, Italy; e-mail: fosca.giannotti@sns.it; Dino Pedreschi, Department of Computer Science, University of Pisa, Pisa, Italy; e-mail: dino.pedreschi@unipi.it.

ACM 1557-7341/2026/3-ART
https://doi.org/10.1145/3802522

offer scalability and computational power. Effective collaboration requires both agents to understand the other's reasoning, strengths, and limitations. In this context, uncertainty and disagreement, which naturally arise from the differing features of human and machine reasoning [81, 140, 157], are not treated as errors to be eliminated but as opportunities for inquiry and mutual refinement; hence, the communication and acknowledgement of uncertainty represent essential aspects of the learning process. The interaction within an HDMS is governed by an "abstention mechanism": a routing function that determines when one agent should defer to the other instead of issuing a final decision. Unlike conventional AI focused solely on accuracy, HDMS integrates human and machine capabilities to enhance overall decision quality, transparency, and reasoning alignment [12, 13]. However, achieving this collaboration poses challenges in designing effective communication and control [79, 142, 150].

This survey provides a structured overview of learning paradigms for HDMSs. The core focus is on uncertainty-driven abstention mechanisms, which determine when an AI should act autonomously and when it must defer to human judgement. This capacity of AI systems to "know what they do not know" is key to optimising overall human–AI performance, boosting reliability, and building mutual trust. We formalise and discuss algorithmic approaches that embed this abstention ability into the machine learning process. The analysis centres on the architectures, learning algorithms, and underlying abstention policies that enable the effective integration of human expertise into the AI decision pipeline. We also explore paradigms that allow humans and machines to collaborate during the learning process through direct interaction, providing a general overview of the possible solutions to move HDMS beyond mere orchestration of human and machine outputs.

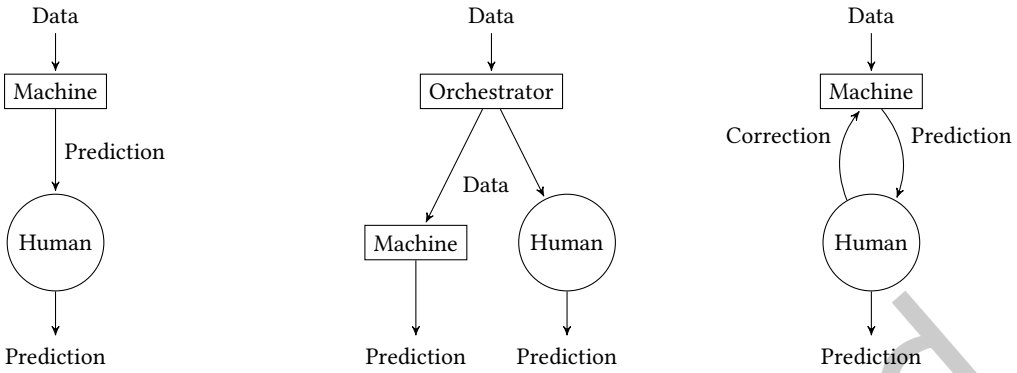
Example of a General Hybrid-Decision-Making System

For clarity, we will use a running example to guide us during the survey. A dermatologist diagnoses skin lesions, diseases, and cancers, e.g., *melanoma*, Detecting melanomas is a binary classification task in which imaging of a skin region is used to detect potentially malignant melanomas. Such a screening task lends itself particularly well to automation, especially in large population screenings. Advancements in AI-based medical imaging have produced models able to often achieve a good enough performance to properly aid health professionals in the field. However, the complexity and the high-stakes nature of the task still demand the dermatologist to play a role. Therefore, what is needed is for the physician to use the AI-based system to enhance their decision-making capabilities in a hybrid decision-making system.

As an outcome of this survey, we propose a taxonomy distinguishing three main learning paradigms for HDMS (Figure 1), reflecting progressively deeper levels of human-machine integration and control over the learning process:

- ***Paradigm 1: Human Overseeing* (HO).** The simplest and most widely used approach. It involves no interaction; the human merely oversees the machine's predictions and may or may not accept the machine's suggestion on the basis of their own knowledge and trust.
- ***Paradigm 2: Learning to Abstain* (LA).** This paradigm uses an algorithmic orchestrator to determine which agent (human or machine) makes each prediction. These methods are the core of this survey, as they provide the methodological basis for formalising abstention policies.
- ***Paradigm 3: Learning together*** (LT). This is the deepest level of integration, establishing a bidirectional communication channel where human and machine engage in an iterative learning loop. Agents exchange information and adapt to one another for a synergistic resolution. Research within this paradigm spans a multitude of very diverse approaches, so rather than providing a systematic review, we discuss representative works and highlight emerging trends and future directions for research on HDMSs.

.



(a) Human overseers. (b) Learning to Abstain. (c) Learning together.

Fig. 1. Paradigms of hybrid systems, where human (circle) and machine (rectangle) collaborate to form a cohesive system. In human oversight (1a) the machine performs a prediction and the human accepts it or rejects it in favour of their own. In Learning to Abstain (1b) an orchestrator assigns the prediction task to either of the two, each of which makes the prediction independently. In Learning Together (1c) the two agents engage in continuous interaction: the machine communicates its reasoning to the human, enabling the latter to understand the machine's internal mechanisms along with possibly rectifying any errors.

Paper selection criteria. Our taxonomy of three families of learning paradigms for hybrid systems was derived from an extensive review of the scientific literature. The paper selection criteria were adapted according to the relevance of each paradigm to the objectives and focus of this study. As for **Paradigm 1**, we have examined papers that have investigated the impact of human oversight on ML models through qualitative or quantitative analysis to highlight and motivate the need for an abstention mechanism in hybrid decision-making. On the contrary, in **Paradigm 2** we conducted a comprehensive review based on the keywords *"learning to reject"*, *"learning to defer"*, *"learning with a reject option"*, *"selective classification"*, *"deferral policy"*, *"deferral function"* and *"defer to expert"*. After a preliminary screening of top conference/journal papers and highly cited papers, where we gathered 150 papers focusing mainly on endowing algorithms with abstention mechanisms, we selected 37 papers representing the wide spectrum of solutions of Learning to Reject and 63 papers encompassing the most significant findings in Learning to Defer. Finally, in **Paradigm 3**, due to the novelty of the topic and the heterogeneous use of the term "hybrid", which also finds wide application in the human-machine interfaces literature, we have gathered journals and papers matching keywords "active learning", "human feedback", "learning with feedback", "interactive learning", "privileged learning", "human AI team", "human in the loop", "training * feedback", and "human * advice", filtering down to a subset of relevant manuscripts, then exploring the cited papers and repeating the process until no relevant manuscripts were found. The chosen papers represent a promising future path for research on hybrid decision-making systems. All papers were searched using *Google Scholar* and the *DBLP computer science bibliography*, selecting papers with high citation counts and/or published in top journals or conferences, such as AAAI, IEEE, ACM, NeurIPS, etc. This work primarily concentrates on *discriminative machine learning models*, which, owing to their robustness, interpretability, and possibility for uncertainty estimation, provide a more suitable foundation for studying abstention mechanisms in HDMSs. While recent advances in generative AI (GenAI) have broadened the landscape of human–AI collaboration, current generative systems remain unreliable for high-stakes decision-making due to their unpredictability and lack of explainability. Large language models (LLMs) further worsen concerns about trustworthiness and safety, as more evidence is emerging pointing to their tendency to replicate and amplify societal biases inherent in training data and their susceptibility

to maliciously designed inputs [73, 154]. Recognising the novelty and appeal of this technology, we believe a dedicated survey on GenAI will be warranted in the future. Nevertheless, we present a concise discussion of how LLMs fit into our paradigms from the perspective of the realisation of an HDMS, highlighting pros and cons of the various technical solutions. For in-depth analysis of base systems, applications in specific domains or tasks, and learning paradigms enabling them, e.g., reinforcement learning, conversational agents, and language modelling, we refer to more in-depth works in the literature [121, 155, 166, 179].

The survey is organised as follows: in Section 2 we provide a general formulation of hybrid systems, setting the stage for the subsequent discussion of our proposed taxonomy. Then, in Sections 3,4, and 5, we survey and discuss the proposals in the literature corresponding to each of the three learning paradigms, namely Human Oversight, Learning to Abstain and Learning Together. In Section 6, we tackle the application of LLMs in HDMSs and their alignment with our proposed taxonomy. Finally, we provide an overview of the open problems and concluding remarks in Section 7.

2 General formulation of Hybrid Decision-Making Systems

A Hybrid System is composed of two types of agents: a machine agent M and a human agent H. Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ such that Ω is a sample space (i.e., an arbitrary non-empty set), $\mathscr{F}$ is an event space (i.e., a σ-algebra over Ω), and $\mathbb{P}$ is a probability measure $\mathbb{F} \rightarrow [0, 1]$. Moreover, let $\mathcal{X}$, $\mathcal{Z}$, $\mathcal{Y}, \mathcal{Y}_M$, and $\mathcal{Y}_H$ be measurable spaces such that $\mathcal{X}$ represents the machine feature space, $\mathcal{Z}$ the human expertise that can be modelled as features, decision rules, etc., and $\mathcal{Y}, \mathcal{Y}_M$, and $\mathcal{Y}_H$ the ground truth, and machine and human label spaces for a certain task T, respectively. The full list of symbols can be found in Table 1.

Machine model. Let $X : \Omega \rightarrow \mathcal{X}$ and $Y_M : \Omega \rightarrow \mathcal{Y}_M$ be random variables representing the input and output of a machine M, and let $Y^* : \Omega \rightarrow \mathcal{Y}$ be the random variable representing the true labels. In the classical setting of supervised learning, given independent and identically distributed pairs $\{(X_i, Y_i^*)\}_{i=1}^n \overset{\text{iid}}{\sim} (X, Y^*)$ drawn from the same unknown joint distribution over $X \times Y^*$, we aim to learn a predictor $f : \mathcal{X} \rightarrow \mathcal{Y}_M$ that approximates, as accurately as possible, the unknown function $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ representing the true relationship between the features and the target. Note that f is chosen from a space of hypothesis $\mathcal{F}_M$ parameterized by $\theta \in \Theta$, however, in our work we often omit the parameter θ to simplify our notation. The ultimate goal of the learning algorithm is to find a hypothesis $f \in \mathcal{F}_M$ that minimizes the *empirical expected risk*, which is defined on the training set as follows:

$$\widehat{\mathcal{R}}_n[f] = \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}_M(Y_{M,i}, f(X_i)) \tag{1}$$

The function $\mathscr{L}_M : \mathcal{F}_M \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_{>0}$ represents the machine loss, which quantifies how far the predictions of a hypothesis $f \in \mathcal{F}_m$ are from the true outcome.

Human model. Let $Z : \Omega \rightarrow \mathcal{Z}$ and $Y_H : \Omega \rightarrow \mathcal{Y}_H$ be random variables representing, respectively, the human expertise and the predictions of a human agent H for a certain task T, where H is modelled as a predictor $h : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}_H$, with h selected from a hypothesis space $\mathcal{F}_H$. Generally, we consider $|\mathcal{X} \bigcap \mathcal{Z}| \geq 0$, implying that there may exist shared information between the machine and the human, or there may not.

It is worth noting that there may be a divergence between the predictions made by human agents and the ground truth labels. Indeed, different levels of background knowledge, experience, or personal biases can lead to distinct decision-making outcomes, resulting in both correct and incorrect predictions across various domains within the input space [88]. Therefore, we additionally take into account a loss function $\mathscr{L}_H : \mathcal{F}_H \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_{>0}$ as an indicator of the quality of human predictions. The observation that various types of errors are not only made by distinct human agents but also occur between humans and AI models provides support for the transition towards paradigms that incorporate hybrid combinations of human and machine predictions [82].

Human-AI gating model Enabling synergistic collaboration in an HDMS requires the definition of a gating function that regulates the transition of control between humans and machines, allowing the system to dynamically determine which agent should act at a given moment. In the literature covered by this paper, this function is also referred to as a rejection, deferral, or abstention policy, to reflect contextual variations (precise definitions will be given in Section 4). Consider now the most simple scenario of an HDMS with two participants, H and M. We can construct a gating function $\rho : \mathcal{X} \to \{0, 1\}$ defined as

$$\rho(X) = \begin{cases} 0 & \text{if } M \text{ will give a prediction for the input } X \\ 1 & \text{if } H \text{ will give a prediction for the input } X \end{cases}$$

Note that the action of switching agent may require additional costs.

Consider now the most simple scenario of an HDMS with two participants, H and M. The training objective of M may be modified to include the potential activation of H by learning both the predictor f and the gating function ρ. Hence, the goal is to find the pair $(\hat{f}, \hat{\rho}_M)$ that minimises the system loss $\mathscr{L}_{HDMS}$ which can be expressed as the weighted linear composition of the machine loss $\mathscr{L}_M$ and the human loss $\mathscr{L}_H$:

$$\mathscr{L}_{HDMS}(f, h, \rho, X, Y^*) := \underbrace{\mathbb{1}_{\rho(X)=0}}_{\text{machine action}} \underbrace{\mathscr{L}_M(f, X, Y^*)}_{\text{machine cost}} + \underbrace{\mathbb{1}_{\rho(X)=1}}_{\text{human action}} \underbrace{\mathscr{L}_H(h, X, Y^*)}_{\text{human cost}} \tag{2}$$

where $\mathbb{1}$ denotes the indicator function, while the individual losses $\mathscr{L}_M$ and $\mathscr{L}_H$ can take several forms to account for different "costs", such as the misprediction error as in the 0-1 loss $\mathscr{L}_{0,1} : (\hat{f}, X, Y^*) \mapsto \mathbb{1}_{[Y^* \neq \hat{f}(X)]}$, or the cost of querying the human agent.

In many cases, the formulation of the loss in (2) may have a level of complexity that renders direct computation intractable or computationally undesirable. For instance, in the typical scenario of multivariate classification, the learning objective is formulated as the minimisation of 0-1 loss $\mathscr{L}_{0,1}$, which is equivalent to minimising the misclassification rate. Even with this straightforward selection of the loss function, optimisation remains complex due to the curse of dimensionality and possible model mis-specification in the case of approximation, but also due to the discontinuity and discrete nature of the 0-1 loss in the case of direct computation, which renders the problem neither continuous nor differentiable, hence extremely hard to optimise and computationally intractable for many nontrivial classes of functions [57, 120]. The approach typically employed to overcome this issue is to define and optimise a *surrogate loss function* $\widetilde{\mathscr{L}}_{HDMS}$, that is, a function with good computational guarantees (e.g., differentiability and convexity) that can be easily optimized and whose optimal values approximate well the minimiser of the original computationally hard loss function [14, 97]. The specific formulation of a surrogate loss function is not straightforward, as it depends on the particular task at hand and the desired properties one seeks to guarantee. Refer to the Appendix for a discussion about the fundamental mathematical properties to consider while assessing the suitability of surrogate losses.

3 Human Oversight

Human oversight [86] (HO) is probably the simplest and most straightforward form of hybrid system. In this first paradigm, machine and human agents are independent of each other, the former performing a task and the latter *verifying* its predictions. Informally, the human agents perform a straightforward task: given the machine computation and/or the input data, either accept or reject the computation. More formally, a *human oversight policy* $\rho : \mathcal{X} \times \mathcal{Y}_M \times \mathcal{Z} \to \{0, 1\}$ is a binary function that, given the prediction $Y_M = f(X)$ of a machine M, an overseeing human H with additional expertise $Z \in \mathcal{Z}$, and some input data $X \in \mathcal{X}$, either *accepts* or *rejects* the prediction Y_M:

	Symbol	Definition	Description
Agent	H	$\mathcal{H}$	Human agent.
	M	$\mathcal{M}$	Machine agent.
R. Variables	X	$\Omega \to \mathcal{X}$	Feature matrix. $\mathcal{X}$ includes the empty element $\emptyset$.
	$Y_{(\cdot)}$	$\Omega \to \mathcal{Y}_{(\cdot)}$	Label vector: $\mathcal{Y}_{(\cdot)} \subset \mathbb{R}$ (regression), $\mathcal{Y}_{(\cdot)} \subset \mathbb{N}$ (classification) provided by a given agent $(\cdot) \in \{M, H\}$.
	Y^*	$\Omega \to \mathcal{Y}$	Ground truth label vector.
Machine	f, f_θ	$\mathcal{X} \to \mathcal{Y}_{(\cdot)}$	Predictor function implemented by a machine agent $(\cdot)$, belongs to the family of functions $\mathcal{F}_M$. The parameter $\theta \in \Theta$ is omitted if not relevant.
	f^*	$\mathcal{X} \to \mathcal{Y}$	Ground truth function. Belongs to the family of functions $\mathcal{F}_M$
	$\mathscr{L}_{(\cdot)}$	$\mathcal{F}_{(\cdot)} \times \mathcal{X} \times \mathcal{Y}_{(\cdot)} \to \mathbb{R}_{>0}$	Real-valued loss function of an agent $(\cdot) \in \{M, H\}$.
	$\widetilde{\mathscr{L}}_{HDMS}$	$\mathcal{F}_M \times \mathcal{F}_H \times \mathcal{R} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{>0}$	Real-valued surrogate loss function of a machine agent M.
	ρ	$\mathcal{X} \to \{0, 1\}$	Abstention policy belongs to the family of functions $\mathcal{R}$.
Human	$Z_{(\cdot)}$	$\mathcal{Z}$	Set of artefact(s) only available to the human agent $(\cdot)$.
	$h_{(\cdot)}$	$\mathcal{X} \times \mathcal{Z} \to \mathcal{Y}_H$	Predictor function implemented by the human agent $(\cdot)$.

Table 1. Table of symbols. We will use lowercase letters for elements of a space, e.g., elements x of $\mathcal{X}$. Moreover, $\mathbb{R}$ represents the set of real numbers, $\mathbb{N}$ represents the set of natural numbers and Ω is the sample space (i.e., an arbitrary non-empty set).

$$.\rho(X, Y_M, Z) = \begin{cases} 0 & \text{accept the machine prediction } Y_M \\ 1 & \text{reject } Y_M \to H \text{ predicts} \end{cases}$$

Example of melanoma detection: Human oversight (HO)

In the paradigm of Human oversight, the dermatologist is presented with an image (i.e., the prediction given by the machine) and has to decide whether to agree or not with the prediction of the machine, i.e., whether the given picture is a melanoma or not.

Other than simply leveraging their own expertise Z, overseers often leverage external factors in their decision. In the simplest of cases, rejected patterns of misbehaviour are limited to **machine-specific** failures in which the underlying context is of little or no impact. In these context-independent scenarios, the overseers aim to identify machine failures induced by machine-specific causes by tracking a set of *subjects of monitoring* around which the overseeing policy will be centred. Machine-specific failures may be induced by wildly different factors in each machine, hence we abstract over the underlying causes, since the goal of machine oversight is to *identify*, rather than *diagnose*, undesired behaviour. We tackle behaviour correction later in Section 5. Common subjects of monitoring are:

- *Data shift.* Data shift is generally intended as a change in the data distribution [132], and may be due to a change in feature distribution, i.e., *covariate shift* or knowledge drift, or label distribution, i.e., *prior shift*; these shifts may occur frequently in real-world scenarios, namely due to data seasonality, sampling bias, or naturally occurring distribution changes. Unlike spontaneous *outlier* instances, dataset shifts are responsible for consistent and predictable failures in the model; thus, they have to be accounted for.
- *(Partial) model performance.* Partially stemming from data shift, model performance is another primary subject of interest. Here, we distinguish between two classes of performance metrics: "global metrics", where the model is evaluated *wholly*, and "partial metrics", where the model is evaluated *partially* on a

suitable subset of data. Data shift may indeed only affect a subset of data; hence, global metrics may easily deceive the overseers.
- *Model uncertainty.* Classical learning algorithms are trained to predict, setting *confidence calibration* aside. Simply put, they operate under a *closed-world assumption* where the only available option is to give a prediction; that is, the model has no notion of uncertainty nor of the unknown. This projects a false high confidence, seldom tricking the algorithm user into overestimating its competency [21].
- *Decision complexity.* Even with highly sophisticated and precise models, some decisions are inherently suitable for human rather than algorithmic reasoning [125]. In this context, it is crucial to anticipate which decision is which.

Operating as self-contained agents, machines often lack the decision context wherein their predictions are evaluated and applied. Failures in this domain are said to be **context-dependent**. Context can be of primary importance, and humans are far better suited than machines in understanding it and integrating it in their decision-making process. For a human, concerns such as fairness, legality, and explainability of the decision are strong contextual motivations that a machine does not necessarily take into proper account. Unsurprisingly, most of them are already being encoded in several legislatures, which strongly discourage or punish discriminative or otherwise illegal [9], unexplainable [106] decisions and behaviors. Much of this stems from the current use of machine agents in ethically-charged contexts. For instance, machines are leveraged in monitoring and discouraging [9] illegal activities, where they often yield unfair or biased predictions [48]; furthermore, they are a critical component of speech regulation, an extremely dynamic use case where human scrutiny and decision autonomy are essential, yet they often regulate marginally- or fully-free [55] speech; they aid hiring in public and private companies, yet they are biased [133].

What's more, context is often dynamic and loosely defined [117], and thus integrating it into the machine is an open challenge in and of itself. Jointly, machine-specific and context-specific failures offer a strong motivation for machine oversight. Yet, even though the *why* is clear, *how* machine oversight is to be implemented is still an open problem. Even worse, machine oversight poses a set of inherently human problems to face.

3.1 Oversight pitfalls

While technical solutions to machine-specific failures have already been developed, context-dependent failures cause a plethora of additional and more complex problems. Given that humans are usually the time bottleneck when it comes to decision-making, one cannot let the overseers monitor every prediction, thus one needs to understand *when* to let the overseer monitor the machine. A conventional solution, which we explore in Section 4, is to let the machine itself call the human into action. Here, we focus instead on the overseers and their inherently human fallacies, which lead to some natural pitfalls of the whole monitoring process.

Human Executors and Skeptics. Overseers need to be aware of two possible cognitive biases: algorithmic *aversion* [41] and *overreliance* [101]. Algorithmic aversion pushes the overseers towards excessively doubting the machine, thus introducing unnecessary monitoring in the decision-making process. Algorithmic aversion manifests itself independently of the performance of the machine [41, 105], and more strongly when the machine fails. In other words, every single perceived mistake of the machine compounds in increasing the rejection rate of the overseer. Unlike algorithm aversion, algorithm overreliance occurs when human agents under-monitor a machine system, and thus act as mere executors. The two biases are well-documented in the literature, and regardless of the machine system they human agents interface with, they are to be accounted for.

Biased oversight. Automation bias is particularly strong when human agents oversee fairness-related tasks where the task directly involves other humans. When monitoring decision on pre-held stereotypes, say on vulnerable groups, overseers either avoid monitoring in the first place [2] or further confirm the stereotypes [10, 48]. On similar reasoning, particularly in cases of fairness evaluation, human agents tend to under-monitor when they perceive affinity towards the monitor case at hand [60], or simply when they deem their reasoning more "human" than the machine's [93]. On an even more biological level, intrinsic demographic traits are also likely to play a role in the decision-making of the human agents [127, 168]. To further increase the complexity of setting up a set of overseers, it is often the case that human agents, in part for the aforementioned reasons, have a low level of agreement on the correct task solution [59].

Failure to oversee and trust calibration. Even more worrying than biased monitoring is the failure to reject obvious machine failures. In a pilot study with legal experts, [45] showed that, when assisted by a faulty machine agent, domain experts incorporate into their decision-making machine recommendations based on irrelevant or random factors, with extreme cases in which the domain experts were knowingly introducing random factors themselves – a clear case of the placebo effect where the mere presence of a machine prediction, regardless of its correctness, induces an almost blind trust in the human agent. Unsurprisingly, overseers have repeatedly been shown to be unable to properly assess their ability to assess the performance of a given machine [1, 152].

Oversight as motivation for abstention Human oversight highlights four recurring problems that make machine-initiated abstention valuable: (i) distribution shift causes unpredictable model errors; (ii) miscalibrated confidence leads users to over- or under-trust predictions; (iii) context- dependent decisions are often invisible to black-box predictors; (iv) human cognitive biases limit scalable and accurate manual review. Abstention directly mitigates (i)–(iii) by enabling automatic triage, and it mitigates (iv) by limiting human load to cases where their added value is maximised.

3.2 Enhanced oversight: Explainable AI

Overseeing a machine simply through its predictions and uncertainty provides minimal tools to a human agent, who can easily fall into one or more of the aforementioned pitfalls. To enhance their overseeing power, humans are often accompanied by *explanation algorithms*, that is, algorithms able to further explain the predictions given by a machine. Explainable AI (XAI) [62] is a recent field of research aiming to shed light on the prediction process of AI models by extracting human-understandable explanations. Explanations allow an overseer to peer into the machine and get a grasp of what features a machine is relying upon [102] and what rule-like logic it is following [61] to make its predictions, what training instances have had a particular influence on the learning process [85], and how one could change the input instance to achieve a different prediction [170]. Explanations have shown to empower the human agent into better understanding the machine, thus improving their ability to monitor it.

4 Learning to abstain

To mitigate human failures in the monitoring process of the aforementioned HDMS paradigm, a potential approach involves developing an enhanced machine architecture that enables the machine learning model to refrain from predicting on certain instances. This is called an "abstention" option: whenever a machine's confidence is insufficient, it incurs a small extra cost to reject the prediction or defer that instance to a human expert [35]. Such Learning to Abstain (LA) systems thus learn, without any fixed interaction rules, to route each input, either predicting themselves or sending it on to another agent. The human decision-making in this paradigm may come afterwards, e.g., operating on those instances rejected by the LA system.

LA systems can be broadly categorised into **Learning to Reject** (L2R) [27, 35] and **Learning to Defer** (L2D) [103, 115] systems, depending on whether the model assumes a fixed cost for abstaining or instead is designed to adapt rejection with respect to a human agent. For both, the setup is the same: the machine learns both a classifier and a *rejection or deferral function* under the optimisation objective of maximising the performance of the human-AI system as a whole. Although rejection and deferral have been tackled as separate and independent problems, a seminal paper by Madras et al. [103] demonstrated that the reject option can be regarded as the specific case of deferral in which a fixed cost is allocated to each deferred instance. Therefore, we devote most of our discussion to L2D (Section 4.2) while giving a more general overview of L2R (Section 4.1).

4.1 Learning to Reject

Learning to Reject (L2R) was first introduced by Chow [27], and its general formulation constitutes a base for more advanced learning to defer algorithms (Section 4.2.3).

Example of melanoma detection: Learning to Reject

In the L2R paradigm, the algorithm is trained to observe the image of a nevus and evaluate whether to make a prediction or abstain from it. For example, edge cases which require further analysis or out-of-distribution images with peculiar melanoma shapes, sizes, and colours, or even images of patients with peculiar skin complexions, e.g., due to skin disease or poor imaging. In such a scenario the dermatologist would again observe the output of the algorithm, that is, a prediction of malignant or benign melanoma, or a rejection of the image.

In the literature, this area of research is referred to with several names: *learning to reject* [182], *selective classification* [43], or *machine learning with a reject option* [70]. Learning to Reject (L2R) equips ML models with the option to abstain on uncertain or difficult instances, improving final performance by adding a reject option. Humans aren't active in this framework, although in hybrid systems rejected cases can be routed to human overseers; L2R has a long research history and is extensively surveyed. Therefore, we report a similar general definition to the work of Hendrickx et al. [70] and Zhang et al. [182]. The goal of L2R algorithms is to learn a model f_ρ composed of two parts: a prediction function $f : \mathcal{X} \to \mathcal{Y}_M$ and a rejection policy $\rho : \mathcal{X} \to \{0, 1\}$ The composed system can be defined as $f_\rho : \mathcal{X} \to \mathcal{Y}_M \bigcup \{\emptyset\}$ such that:

$$f_\rho(x) = \begin{cases} \emptyset & \text{if } \rho(x) = 1 \\ f(x) & \text{otherwise} \end{cases} \tag{3}$$

That is, if the rejection policy ρ rejects x, then no prediction is made. If instead ρ accepts x, then the prediction function f is applied to x and the result $f(x)$ is obtained. The function f is generally assumed to be a classifier. Ideally, ρ should be able to prevent misprediction of f while conversely accepting instances for which a good prediction is more likely. The policy ρ that actually performs the rejection operation is called *rejection function* or *rejector*. The focus of a rejector, as mentioned, is to refuse those instances for which the classifier f is expected to output an incorrect prediction.

More broadly, research in this field seeks an optimal trade-off between accuracy on accepted instances and the number of rejections. Rejections may target either *novelties*, instances far from the training set X, i.e. out-of-distribution or unusual instances, or *ambiguities*, instances near the decision boundary whose class-membership probabilities are nearly equal [64].

The rejector ρ can rely solely on the feature space $\mathcal{X}$ or also on the predictor f's output or internals. In the former *independent* setting, ρ is learned agnostic to f; in the latter *dependent* setting, ρ is constructed by querying or exploiting properties of f.

4.1.1 Independent Rejectors. Independent rejectors mainly address novelty rejection by treating OOD instances as outliers or anomalies [71, 160], or under the open-set recognition paradigm [104], flagging them directly from the data without reference to f.

Early approaches leveraged statistical anomaly detection: Seo et al. used Gaussian process posterior variance to reject outliers [146], and Coles applied Extreme Value Theory models—later used in facial recognition [144] and network analysis [110] [34]. More recently, Rudd et al. introduced kernel-free, variable-bandwidth incremental learning via Extreme Value Machine. Although this method may reject classes for their distance with respect to the majority, this is mitigated by using a generalized Pareto approximation for anomaly thresholds [164].

Alternatively, some methods train anomaly detectors on the in-distribution data: One-Class SVMs [32], Gaussian Mixture Models [89], and few-shot margin-loss models for novel-class detection [5]. Beyond pure anomaly rejection, Asif and Amir Afsar Minhas proposed jointly trained neural networks with dual penalties on misprediction and rejection—applicable to both novelty and ambiguity rejection—while remaining model-agnostic [7].

4.1.2 Dependent rejectors. Dependent rejectors, which exploit the predictor's outputs or internal characteristics, are classified as either *staged* or *joint* (cf. Section 4.2).

Staged rejectors. These rejectors estimate model confidence or uncertainty after f is trained. Formally, $\rho(x, f) = \mathbb{1}_{c(x,f)>\tau}$, where $c(x, f)$ is a confidence metric and τ a threshold [43, 65]. Metrics based on *hard predictions* $y_M = f(x)$ assess class-wise variance over repeated predictions [16, 151], while *soft predictions* use scoring outputs approximating $P(y_M \mid x)$ [19, 40]. Other approaches use the SVM decision-boundary score [161] or the distance to the k-th nearest prototype [20]. The threshold τ may be *global*, for uniformly calibrated models [24, 50, 90], or *local*, multiple τ_i, for variable accuracy regions or class-wise variance [49, 128].

Joint rejectors. Also called integrated rejectors [70], these rejectors treat rejection as an additional class learnt alongside the predictive classes, making f and ρ indistinguishable [35]. Many methods minimise a single objective that penalises errors and rejections via surrogate losses [15, 137]. Others optimise specific metrics, e.g., AUC for binary classifiers with rejection [131], or add a rejection class with its own cost [186]. Model-specific joint learners have been developed for SVMs [58, 96] and neural networks [54].

4.1.3 Cost model for rejection. Approaches for L2R must trade off predictive performance against rejection rate. In the absence of an appropriate cost model, a classifier with a reject option can either reject everything to eliminate errors or reject nothing to maximise coverage, so a cost model is required to prevent these extremes. One foundational cost model can be found in the work of Cortes et al.[35]:

$$cost(x) = \begin{cases} 0, & \text{if } \rho(x) = 0 \wedge f(x) = y^* \\ 1, & \text{if } \rho(x) = 0 \wedge f(x) \neq y^* \\ \mathcal{R} & \text{if } \rho(x) = 1. \end{cases} \tag{4}$$

where y^* is the ground truth, for the instance x, and $\mathcal{R}$ is a fixed, predefined cost for rejection. If we adopt such a cost model, we can express the learning objective of an L2R algorithm as follows:

$$\min_{f,\rho} \sum_{x \in X} [\, \mathbb{1}_{\rho(x)=0} \mathbb{1}_{f(x)\neq y^*} + \mathbb{1}_{\rho(x)=1} \mathcal{R}] \tag{5}$$

We incur a cost when mispredicting ($\mathbb{1}_{f(x)\neq y^*}$) an accepted instance ($\rho(x) = 0$) or a cost $\mathcal{R}$ when rejecting it ($\rho(x)\mathcal{R}$). While sensible, this approach requires predefining $\mathcal{R}$ which may not always be feasible [54]. Geifman and El-Yaniv addressed this by introducing Selection with Guaranteed Risk Control, focusing on model coverage, defined as the fraction of data accepted for prediction: $\frac{1}{n} \sum_{x \in X} \mathbb{1}_{\rho(x)=0}$. Their objective minimizes risk while ensuring the model predicts a minimum coverage. Formally:

$$\min_{f,\rho} \frac{\sum_{x\in X} \mathbb{1}_{\rho(x)=0}\mathbb{1}_{f(x)\neq y^*}}{\sum_{x\in X}(1-\rho(x))} \quad \text{s.t.} \frac{1}{n}\sum_{x\in X} \mathbb{1}_{\rho(x)=0} > C \tag{6}$$

where $0 < C < 1$ is a coverage threshold. The problem has thus shifted from having to determine the cost $\mathcal{R}$ to the simpler task of selecting a threshold C. It is important to note that if $C = 1$ the model reverts to a standard classifier with no rejection option. This coverage-based formulation has been proven to be theoretically equivalent to the original cost model of Eq. (4) [51].

4.1.4 Strengths and limitations of Learning to Reject. In summary, the L2R paradigm allows for the development of ML models with a reject option, which is a foundational starting point for developing models able to interact with humans. Indeed, ML models equipped with the reject option can, in principle, reject exactly those instances that would yield a prediction error and therefore call for human intervention only when strictly needed. However, the actual benefit of these techniques in a collaborative setting with humans has never been thoroughly investigated. Indeed, if human intervention is called only for those instances for which a decision is difficult, the human expert may find the same difficulties and thus deem the model as not so useful for solving the task. Moreover, there are studies pointing to possible fairness issues when using classifiers with a reject option [76].

4.2 Learning to Defer

Learning to Defer (L2D) systems embed human knowledge directly into the training process of a ML model. The goal is to equip the model with the ability to call for human intervention in those instances where the human is likely to give an accurate prediction and the machine is likely to fail. In contrast to L2D, an L2R system learns its rejector policy only from the feature set $\mathcal{X}$ (e.g., the text of the message to be flagged) and, possibly, some properties of the predictor used by the AI system. L2D instead actively considers the human expertise in the task domain. In this section, we present the state-of-the-art literature on L2D based on the selected works summarized in Table 2.

Example of melanoma detection: Learning to Defer

By comparing predictions made by the dermatologist on previous cases to the correct labels, an L2D system can be trained to determine which instances can be accurately predicted by AI and which are better handled by (a committee of) humans. For instance, in the case of melanoma detection, humans are expected to outperform machines on dubious or extremely difficult cases, mostly owing to their greater capacity for comprehending common sense and contextual information, as well as integrating external factors regarding the patient, such as for example, different skin complexions and other clinical data.

4.2.1 General formulation. Keeping the same notation introduced in Section 2, we consider HDMS composed of a machine M and a human H. Similarly to L2R, in L2D the machine M is equipped with the possibility of abstaining from making a prediction. In addition, an L2D model also embeds a representation of the human agent H, thereby taking into account their estimated performance when assessing the act of deferral. By doing so, the human expertise $\mathcal{Z}$ is taken into account. Nevertheless, note that the predictor $h : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}_H$ modelling H is *fixed*, meaning that L2D algorithms have no control nor visibility over the function h itself; rather, they only have access to its image, that is, the set $\{Y_{H,i}\}_{i=1}^n \coloneqq \{h(x_i, z_i)\}_{i=1}^n$ of human predictions about the training data.

In analogy to Eq. (3), a L2D system can be formulated as a function $f_\rho : \mathcal{X} \to \mathcal{Y}_M \bigcup \mathcal{Y}_H$ defined from the classifier $f : \mathcal{X} \to \mathcal{Y}_M$ and deferral policy $\rho : \mathcal{X} \to \{0, 1\}$:

$$f_\rho(x) = \begin{cases} h(x) & \text{if } \rho(x) = 1 \\ f(x) & \text{otherwise} \end{cases}$$

Architecture	*Design according to which classifier and deferral policy are integrated.*	
Staged learning	Classifier and deferral policy are learnt in separate steps.	[11, 134, 135, 174]
Joint learning	Classifier and deferral policy are learnt jointly.	[25, 52, 67, 83, 98, 103, 114, 115, 130, 136, 162, 163, 174]
Others	Alternative solutions, e.g., iterative models.	[38, 39, 122]
Multiplicity	*Number of human agents in the hybrid system.*	
Single	One human agent.	[11, 25, 98, 103, 114, 115, 122, 130, 134–136, 163, 174]
Multiple (1 predicts)	One human is selected out of many.	[52, 67]
Multiple (j predicts)	A subset of agents is selected out of many.	[83, 162]
Theoretical guarantees	*Theoretical guarantees of the model.*	
Fisher consistency	The optimisation objective has the correct target.	[25, 115, 162, 163]
Classification-calibration	Agents are given realistic uncertainty estimates.	[162, 163]
realisable consistency	The problem is well-defined under specific choices of the classifier and deferral hypothesis function spaces.	[114]
Constraints	*Additional conditions that the hybrid system should satisfy.*	
Coverage	Number of instances that can be deferred.	[39, 114, 115, 122]
Budget	Total cost to query human agents.	[134]
Fairness	Metrics to guarantee algorithmic fairness.	[83, 103, 114]
Others	Others, e.g., on the selection of human agents.	[83]

Table 2. Properties of systems in the Learning to Defer paradigm.

The optimisation goal of an L2D system has the same formulation as Eq. (2), which is a weighted linear combination of the machine loss $\mathscr{L}_M$ and the human loss $\mathscr{L}_H$. However, note that the learnable parameters in this scenario are only f_θ and ρ, since both $Y_H = h(X)$ and Y^* are fixed (i.e., they belong to the training data). In general, the individual losses $\mathscr{L}_M$ and $\mathscr{L}_H$ can take several forms to account for different "costs", such as the misprediction error as in the 0-1 loss or the cost of querying the human agent. When there exists a constant $\mathcal{R} > 0$ such that $\mathscr{L}_H(y^*, y_H) = \mathcal{R}$ for all $(y^*, y_H) \in (Y^*, Y_H)$, then the loss in Eq. (2) matches the rejection loss formulated in Eq. (5) under the assumption of the 0/1 cost model for prediction/rejection [103]:

$$\mathscr{L}_{\text{reject}}(f, \rho, X, Y^*) \coloneqq \mathbb{1}_{\rho(X)=0}\mathscr{L}_M(f, X, Y^*) + \mathcal{R}\mathbb{1}_{\rho(X)=1} \tag{7}$$

Optimisation constraints. Depending on the specific context of use, the application of specific constraints may be necessary for hybrid systems. This objective is commonly accomplished by incorporating regularisation terms into the system loss or by imposing specific bounding conditions. Examples of such constraints include:

- Coverage or triage level [39, 114, 115, 122]: the number of instances that can be deferred.
- Fairness metrics [83, 103, 114]: for instance, the Minimax Pareto Fairness criterion [111] or the equalised odds metric with respect to a protected attribute.
- Budget [134]: total cost that can be allocated to query human agents.

Model architectures. L2D systems typically adhere to either of two general designs, referred to as *staged learning* and *joint learning*, which vary in terms of when the classifier and rejector are learnt. In the former case, the algorithmic process starts by learning the classifier and only subsequently fits the deferral policy on top of it. On the other hand, in a joint learning setting, the classifier and rejector are learnt simultaneously through the direct minimisation of the system loss in Eq. (2) While most of the proposals documented in the literature can be categorised as staged or joint learning models, a few exceptions also exist that do not fit in either category

(Section 4.2.4).

Number of human agents. L2D systems can be characterised along another dimension, that is, the size of the pool of human agents to which the decision can be deferred. Whenever the number of these is greater than one, term *Multiple-Expert L2D* (L2D-ME) is used, as opposed to to *Single-Expert L2D* (L2D-SE or L2D), which considers one human only. An example where L2D-ME modelling may be more suitable is in a medical setting, where a critical decision regarding a complex case could be made either by an automated classifier or by one or more doctors chosen from a team of experts with potentially diverse expertise and opinions. As compared to L2D-SE, the formalisation of L2D-ME makes the deferral function more complex in nature. Specifically, it should not only determine *when* to defer but also to *which* human agent(s) [162]. Furthermore, the deferral policy can be designed in a manner that allocates predictions to either *one* single agent or to a *subset* of agents from the available pool.

4.2.2 Staged learning architectures. In this family of L2D architectures, the classifier f and deferral function ρ are learnt separately: algorithms first fit a classifier on the training dataset, then they learn a second model that predicts the probability that the human makes a mistake on the same dataset, and finally they defer based on which has the lowest error probability instance-wise. For instance, Raghu et al. [134] developed a basic heuristic for L2D consisting of two independent models trained on the full dataset: a multiclass classifier representing the machine agent and a binary classifier representing the correctness of the human agent. At inference time, an instance is deferred to the human if the predicted classifier error probability is higher than that of the human. In case of coverage constraints, then the samples whose difference between human and classifier error probability is higher are chosen first. Interestingly, the authors also suggest a reduction of L2D-ME to L2D-SE by modelling the human subsystem in terms of average disagreement between human agents on each single prediction. This approach has been further developed in [135].

Another common baseline for staged learning is the model proposed by Bansal et al. [11], who described a staged learning setting aimed at maximising the expected utility of the system, which is measured in terms of the accuracy of the final decision, the cost of deferring, and the individual accuracy of both the human and machine components. Differently from other L2D models, this method has been claimed to be user-initiated, since the action of deferral is triggered through an (over-simplified) threshold-based policy that represents the humans' mental model of the AI.

Finally, a third relevant staged learning method known as the *fixed value of information approach* has been proposed in [174]. It consists in training independently three probabilistic models describing, respectively, the distribution of the label given the input data, the human predictions given the input data, and the label given both the input data and human predictions. At inference, the deferral policy evaluates the estimated expected utility of the classifier in two scenarios: when the human is not consulted and when the human is queried, while also taking into account the distribution of human predictions and a constant cost for querying the human.

As noted by Charusaie et al. [25], the staged learning approach presents some important advantages: first of all, it is suited for convenient implementation, since already known appropriate algorithms can be adopted to solve the two stages separately. Secondly, theoretical and experimental results suggest that it outperforms the joint learning approach in realistic scenarios where only a limited portion of data is labelled by the human agent. In these cases, the classifier f can still be optimized over the full dataset; on the other hand, in joint learning f can be learnt from the subset of human-labelled data only, thus leading to a reduction in performance dependent on the proportion of unlabelled data. However, [25] also pointed out that staged learning is sub-optimal with respect to joint learning and provides both theoretical and experimental results showing the existence of a performance gap between the two approaches in terms of model complexity.

4.2.3 Joint learning architectures. In L2D systems characterised by a joint learning architecture, the classifier f and deferral function ρ are learnt simultaneously. In order to implement this design, the task is shaped as a $K + 1$

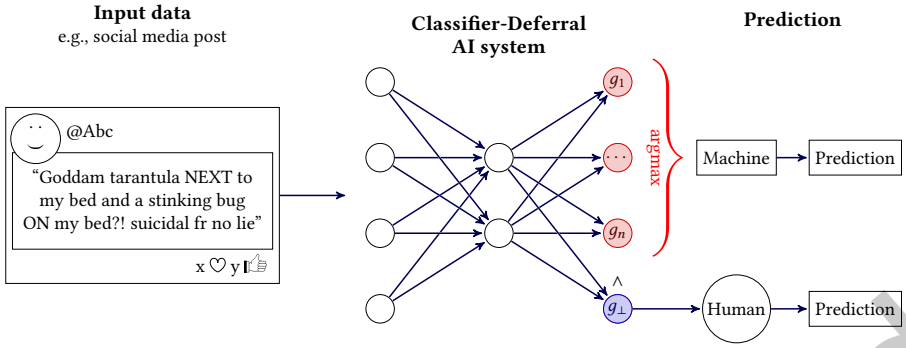


Fig. 2. Overview of the joint learning architecture for the Single-Expert Learning to Defer (L2D-SE) setting, illustrated in the application of flagging online content for moderation. Adapted from [114].

multiclass problem over an augmented label space $\mathcal{Y}^{\emptyset} \coloneqq \mathcal{Y}_M \cup \{\emptyset\}$, where $\mathcal{Y}_M = \{1, \dots, K\}$. In particular, we set $\mathbf{g} = (g_1, \dots, g_K, g_{\emptyset})$ to be the set of real-valued scoring functions $g_i : \mathcal{X} \to \mathbb{R}$, such that $g_{\emptyset}$ returns the human predictions Y_H, while the classifier and rejector are defined, respectively, as:

$$f(x) = \arg\max_{i \in \mathcal{Y}_M} g_i(x) \qquad \rho(x) = \begin{cases} 1 & \text{if } \max_{i \in \mathcal{Y}_M} g_i(x) \leq g_{\emptyset}(x) \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In general, the optimal classifier-rejector pair is found by minimising the system loss expressed in Eq. (2) However, Eq. (2) is often computationally hard to optimise. Such a problem is addressed by replacing Eq. (2) with a *surrogate loss* $\widetilde{\mathscr{L}}_{\text{defer}}$ that is easy to optimise and is chosen to guarantee specific properties with respect to the original loss $\mathscr{L}_{\text{defer}}$ (refer to the Appendix for a formal discussion of desirable properties of surrogate loss functions in the context of L2D). Hence, in the case of appropriate choices of the human and machine surrogate loss functions, the joint learning HDMS gives theoretical guarantees for optimal performance.

Single-Expert L2D. Most of the literature on joint learning models for L2D-SE focuses on the *cost-sensitive formulation* of the problem over an augmented label space $\mathcal{Y}^{\emptyset}$ developed by Mozannar and Sontag [115] and illustrated in Figure 2. This setting considers random costs $\mathbf{c} = \{c_1, \dots, c_{K+1}\} \in \mathbb{R}^{K+1}_{>0}$ where each component c_i represents the cost of predicting the label $i \in \mathcal{Y}^{\emptyset}$. In this section, we review and categorise the most relevant proposals according to their statistical properties.

Fisher Consistency (FC) has been described as a minimal requirement that surrogate loss functions should satisfy to achieve reasonable performance, since it posits that if an estimator were computed using the complete population instead of a sample, it would yield the true value of the estimated parameter [97]. To the best of our knowledge, FC approximations of the 0-1 loss in the L2D setting have been implemented (up to adaptations) only by Charusaie et al. [25], Mozannar and Sontag [115], Verma and Nalisnick [163]. In particular, the surrogate loss $\mathscr{L}^{\alpha}_{CE}$ [115] consists of a generalisation of the cross-entropy loss with the costs corresponding to multiclass misclassification, where the cost of the $K+1$ class represents the act of deferral and $\alpha \in \mathbb{R}_{>0}$ is a weighting parameter that modulates deferral. When $\alpha = 1$, $\mathscr{L}^{\alpha}_{CE}$ has FC and can be expressed as follows:

$$\mathscr{L}^{1}_{CE}(\mathbf{g}; X, Y_H, Y^*) \coloneqq -\log\left(\frac{\exp(g_{Y^*}(X))}{\sum_{y \in \mathcal{Y}^{\emptyset}} \exp(g_y(X))}\right) - \mathbb{1}_{Y_H = Y^*} \log\left(\frac{\exp(g_{\emptyset}(X))}{\sum_{y \in \mathcal{Y}^{\emptyset}} \exp(g_y(X))}\right) \tag{9}$$

intuitively, the first term maximises the scoring function associated with the true label, while the second maximises the rejection (scoring) function but only if the human's prediction is correct. Notably, Charusaie et al. [25] presents

a unified framework that allows the use of any consistent multiclass loss for constructing a consistent surrogate for L2D, thus generalising prior work [23, 115, 163].

A few adaptations have been proposed to enhance the L2D algorithms based on the surrogate $\mathscr{L}_{CE}$, with the aim of better capturing specific properties. These include:

- *Learning to Defer with Uncertainty* (LDU), where the deferral policy accounts for the epistemic uncertainty of the model (i.e., the uncertainty resulting from limited data availability and lack of knowledge about the system of interest) [98];
- The customisation of the model to suit the expertise of a particular human agent, which, however, requires the availability of supplementary data that has been annotated by that particular human [136].
- *Label-smoothing-free loss* (LSF). Narasimhan et al. [118] demonstrated that consistent loss functions experience *underfitting* when the additional cost of deferring to the expert is non-zero, as this scenario introduces a label smoothing term, which results in a flattened training distribution. Liu et al. [99] propose a novel loss formulation which tackles this issue while preserving statistical consistency. Moreover, the authors demonstrate that current representative surrogate losses for L2D [23, 115, 163] can be devoid of label smoothing by plugging their base multiclass losses into their suggested loss formulation, which is as follows:

$$\mathscr{L}_{\psi}^{LSF}(\mathbf{g}; X, Y_H, Y^*) = \psi(\mathbf{g}(X), Y^*)) + c\,\mathbb{1}_{Y_H \neq Y^*} \min_{y \in \mathcal{Y}_M} (\mathbf{g}(X), y) + (1-c)\mathbb{1}_{Y_H = Y^*}\psi(\mathbf{g}(X), K+1))$$

Confidence calibration refers to the property of an estimator (e.g., a probabilistic classifier) to produce a predictive distribution that is consistent with the empirical frequencies observed from realised outcomes [37]. Verma and Nalisnick [163] proposed a surrogate loss $\mathscr{L}_{OvA}$ [163] that satisfies both Fisher consistency and classification-calibration. This solution consists in solving the L2D problem via a One-vs-All classification method which breaks down the original $K+1$ classes into $K+1$ binary classifier models. The resulting objective function to be optimized is thus a surrogate loss function composed of different logistic loss components, each accounting for the error on one of the $K+1$ different classes.

$$\mathscr{L}_{OvA}(\mathbf{g}; Y_H, X, Y^*) \coloneqq \phi[g_{Y^*}(X)] + \sum_{\substack{y \in \mathcal{Y}_M \\ y \neq y^*}} \phi[-g_y(X)] + \phi[-g_\emptyset(X)] + \mathbb{1}_{Y_H = Y^*}(\phi[g_\emptyset(X)] - \phi[-g_\emptyset(X)])$$

where $\phi : \{0,1\} \times \mathbb{R} \to \mathbb{R}_{>0}$ is a binary surrogate loss (e.g., the logistic loss). Experimental findings show that $\mathscr{L}_{OvA}$ results in better calibrated models w.r.t. ones trained with $\mathscr{L}_{CE}$, with competitive performance w.r.t. other L2D baselines [11, 115, 122, 134].

A *bounded* function f is a function defined on a bounded set X of real or complex values, meaning there exists a real integer N such that $|f(x)| \leq N \; \forall x \in X$. While exploring the confidence calibration properties of $\mathscr{L}_{CE}$, Verma and Nalisnick [163] found that this loss is also unbounded, as it can exceed values larger than one, hence failing to adequately calibrate the expert correctness. To tackle this issue, Cao et al. [23] offer a statistically consistent *asymmetric*[1] softmax-based surrogate loss that generates reliable estimates without the miscalibration and unboundedness issues that characterise $\mathscr{L}_{CE}$. It has the following formulation:

$$\mathscr{L}_{\psi}^{ASM}(\mathbf{g}; Y_H, X, Y^*) = -\log(\psi(g(X))) - \mathbb{1}_{Y_H \neq Y^*}\log(1 - \psi^{K+1}(X)) - \mathbb{1}_{Y_H = Y^*}\log(\psi^{K+1}(X)).$$

where ψ is an asymmetric softmax function obtained from the standard softmax function by adding an asymmetry w.r.t. the additional $k+1$ class of the augmented label space. Moreover, the authors also discuss the possibility of extending $\mathscr{L}_{\psi}^{ASM}$ to the multi-expert L2D setting.

A different formulation of the bounded loss function has been given by Wei et al. [172] while studying the

[1]An *asymmetric* learning setting here refers to one characterised by unequal misclassification costs or training data imbalance [145]

possibility of adopting an alternative Bayes optimality definition, that is, the minimiser of the Bayes risk for which any surrogate loss function is demonstrated to be consistent. Differently from the one proposed in current representative L2D models [115], this definition accounts for dependence patterns between humans and models. Specifically, it introduces a novel deferral principle that assesses deferral according to the dependence pattern identified in training data, hence bypassing the necessity for confidence estimation. Motivated by their formulation of *dependent Bayes optimality*, Wei et al. [172] also present a novel *Dependent Cross-Entropy* (DCE) loss that is consistent and capable of inducing a bounded confidence estimator for the expert.

$(\mathcal{F}_M, \rho)$- *Realisable consistency* is a property that refines the notion of FC by addressing the optimisation process over restricted hypothesis classes $\mathcal{F}_M$ and ρ for the predictor and deferral function. For instance, the surrogate $\mathscr{L}_{RS}$ [114] is differentiable, non-convex, and realisable $(\mathcal{F}_M, \mathrm{P}_M)$-consistent for classes $\mathcal{F}_M$ and P_M closed under scaling:

$$\mathscr{L}_{RS}(\mathbf{g}; Y_H, X, Y^*) \coloneqq -2 \log \left(\frac{\exp(g_{Y^*}(X)) + \mathbb{1}_{Y_H = Y^*} \exp(g_\emptyset(X))}{\sum_{\in \mathcal{Y}^\emptyset} \exp(g_y(X))} \right)$$

Other formulations of joint learning L2D-SE also exist. This is the case of:

- the seminal paper by Madras et al. [103], which presents a framework for addressing the L2D problem using a Mixture-of-Experts (MoE) approach, with the deferral policy acting as a gating function. The classifier and deferral function are learnt together by negative log-likelihood minimisation over the augmented label space $\mathcal{Y}^\emptyset$. An alternative version of the algorithm is also introduced, wherein a regularisation component is added to the system loss to account for fairness. Unfortunately, this method was proven to not have FC [115].
- *Preferential MoE* [130], a variant of [103] where human knowledge is encoded in the form of decision rules that should be followed as much as possible, that is, whenever they are applicable and do not decrease the system performance. The algorithm first checks the applicability of the available rules and, in case of a positive response, a deferral function selects whether to rely on the human or machine prediction based on their performance. Notably, the deferral function is chosen to be interpretable (e.g., a linear classifier or decision tree), guaranteeing transparency in the selection of the human agent and also highlighting reasons for forgoing the human-based rules.
- In the *joint value of information method* [174], the three probabilistic models already introduced in the *fixed value of information method* described in Section 4.2.2 are trained together through a single neural network which includes a final Platt calibration layer that guarantees the estimation of meaningful expected utilities. Experimental findings show that joint learning yields greater advantages compared to the analogous staged learning method.
- The *Mixed Integer Linear Program* (MILP) [114] is a scheme to exactly minimise the misclassification error of the HDMS. It comes with generalisation bounds and allows you to provably and easily integrate any linear constraints on the variables. However, it suffers from two limitations: it is computationally expensive and it does not generalise to non-linear predictors.

Multiple-Expert L2D (1 out of J). In this first scenario of L2D-ME, the goal of the multi-expert deferral policy ρ_M^{ME} is to choose either the classifier or *exactly one* human agent from the set of J available ones. Hence, ρ_M^{ME} takes the form of $\rho_M^{ME} : \mathcal{X} \to \{0, 1, \ldots, J\}$, where $\rho_M^{ME}(x) = 0$ means that the classifier decides, while $\rho_M^{ME}(x) = j$ for $j \neq 0$ indicates that the decision is deferred to the j^{th} human agent.Hemmer et al. [67] adopt a mixture of experts (MoE) approach with the deferral policy serving as a gating function that assigns each instance either to the predictor or one specific human agent. The joint learning of the classifier and deferral function is carried out through a surrogate loss function based on the negative log-likelihood of the system. However, subsequent work [162] has proven this surrogate to be not FC and proposed instead two surrogate loss functions, namely one based on cross-entropy and one on the One-vs-All classification, which are consistent with the 0-1 loss in the L2D-ME

setting and extend their single-expert analogue. The experimental findings indicate that the OvA-trained model frequently achieves superior performance compared to both the cross-entropy variant and the MoE baseline [67]. Additionally, it exhibits better calibration in terms of the correctness of agents' decisions. The properties of realisable consistency with respect to a certain hypothesis space have been investigated within the context of L2D-ME Mao et al. [107, 108] as well. In particular, the authors present novel families of surrogate losses that are underpinned by *realisable consistency bounds* [8], indicating the existence of upper bounds on the target estimation loss formulated in relation to the surrogate estimation loss, rendering them more advantageous as they are hypothesis set-specific and non-asymptotic.

Finally, Gao et al. [52] study the problem of L2D-ME in a *bandit feedback* setting (i.e., a sequential dynamic allocation problem). Specifically, the deferral policy is specifically learnt using a supervised learning model that has been previously trained on historical data that reflect human decisions and corresponding outcomes in order to maximise the complementarity of the machine and human agents. However, by doing so, it is assumed that the human agents who generated the historical data are the same individuals who will be assigned decisions at inference time.

Multiple-Expert L2D (j out of J). In the second scenario of L2D-ME, the multi-expert deferral policy is defined as $\rho_M^{ME} : \mathcal{X} \rightarrow \{0,1\}^{J+1}$. For each input $x \in \mathcal{X}$, the goal is to choose the committee of agents $C(x) \subseteq \{0,\ldots,J\}$, possibly including the classifier, who are likely to make the most accurate decision for x. Hence, the i^{th} vector component of the deferral policy will be defined as $\rho_M^{ME}(x)_{(i)} = 1$ for all $i \in C(x)$, and $\rho_M^{ME}(x)_{(i)} = 0$ for all $i \notin C(x)$. In the event that the designated committee comprises multiple agents, the resulting outcome will be an aggregated decision. This setting has been firstly addressed by Keswani et al. [83], who proposed a joint loss function obtained by linearly combining the losses associated with the classifier and deferral function via context-dependent hyperparameters. The authors proved that the combined loss is convex with respect to the classifier and deferral function whenever the loss associated with the former is convex; under such an assumption, it can be optimized using the projected-gradient descent algorithm. Additionally, the authors outlined a few adaptations of the L2D-ME framework to account for potential real-world constraints and requirements:

- *Fair learning*: this variant takes into account the possibility of performance discrepancies that may occur with respect to individuals belonging to different protected categories.
- *Sparse Committee Selection*: this variant enables the deferral function to exclusively choose a limited number of agents on a per-instance basis.
- *Dropout*: this variant aims to reduce the dependence on a single agent and achieve a more equitable distribution of workload.
- *Regularised versions*: additional constraints can be added to the joint framework as regularisers of the loss function. For instance, this solution can be employed in cases where specific costs associated with individual human agent consultations are provided.

Subsequent work [83] further developed this setting to adapt to a *closed* deferral pipeline, wherein the human agents of the HDMS also provided the training labels. This is achieved through an online framework in which input samples are received in a continuous stream. After each prediction is made, which involves aggregating the outputs of agents in the chosen committee, the samples are utilised to retrain the classifier and deferral function.

Alternatively, Verma et al. [162] suggest using Conformal Inference [148] to find ensembles of agents $C(x)$ that include the best agent with high marginal probability. The size of $C(x)$ is computed dynamically as a function of the input x, thereby ensuring optimal utilisation of agent queries. The authors propose two test statistics for the estimation of $C(x)$: a naive score function that sums up the correctness scores of all agents who correctly predict the given instance and a regularised statistic that employs conformal risk control [6] to increase the robustness to noise. The experimental findings demonstrate that the latter approach yields a nearly flawless identification of

the appropriate number of agents. Moreover, the conformal approach exhibits superior performance in system accuracy compared to a fixed-size ensemble of agents.

4.2.4 Further model architectures. While most of the proposals documented in the literature can be categorised as staged or joint learning models, a few exceptions also exist. A notable example often used as a baseline in the L2D literature is the method proposed by Okati et al. [122], namely, an iterative algorithm that optimises the classifier and triage policy alternately. At each iteration, the optimisation process is carried out for the classifier on instances where it outperforms the human agents, while the remaining data points are optimized for the triage policy. The authors show that their method converges to a local minimum. Nevertheless, subsequent experimental studies have shown that this method exhibits lower performance in comparison to other L2D algorithms [114]. Additionally, similar algorithms have been implemented to address the issue of L2D for model-specific settings, namely Support Vector Machines [39] and Ridge Regression [38]. A comprehensive framework for L2D in regression tasks with theoretical guarantees was provided by Mao et al. [109], covering both staged and joint model architectures. Yannis et al. [177] recently introduced a staged L2D approach for multi-task settings (classification and regression) that integrates expertise from multiple experts, ensuring Bayes consistency and realisable consistency for any surrogate with consistent bounds. A key limitation of L2D is assuming that the human expert available at test time matches the one who provided the training data, which is rarely true in practice [94]. Instead, a more realistic approach considers that all potential experts share decision-making similarities. Based on this, Tailor et al. [156] proposed the *Learning to Defer to a Population* framework, an L2D-SE system capable of deferring to unobserved human predictions during training within a defined population was proposed. During testing, it assumes an expert is selected from this population, requiring the L2D system to make deferral decisions despite uncertainty about specific expert behaviour, using meta-learning on a limited context set representing expert capabilities.

4.2.5 L2D with limited human predictions. A significant drawback of L2D is the requirement of human predictions, alongside ground truth labels, for every instance within the training set [94]. Ideally, the L2D system has to be trained on human labels belonging to the same human that will then interact with the system itself. By doing so, the L2D system will learn to complement that specific human [68]. Due to the significant computational and human costs, it is likely that the implementation of the L2D algorithm would be impractical for most real-world scenarios. To address implementing L2D-SE algorithms with limited human predictions, Charusaie et al. [25] proposed *Disagreement on Disagreements* (DoD), an active learning scheme for training a classifier-rejector pair with minimal human queries. DoD operates in two phases: (i) a standard active learning algorithm (e.g., CAL [33]) identifies predictor disagreement sets, and humans are queried on these instances to learn their error boundary; (ii) a consistent classifier-rejector pair is then learnt from pseudo-labelled data. Alternatively, Hemmer et al. [68] presented a three-step approach using limited human predictions to generate synthetic labels. First, an *embedding model* maps instances into feature representations. Next, an *expertise predictor model* approximates human capabilities using semi-supervised learning. Finally, the model produces synthetic predictions for unlabelled instances, usable in L2D algorithms. Empirical results show that few human predictions per class suffice for effective synthetic generation. In the multi-expert L2D context, Alves et al. [3] have tackled the challenge of limited human data availability by introducing the *Deferral under Cost and Capacity constraints Framework* (DeCCaF). This innovative L2D-ME approach utilises supervised learning to estimate the likelihood of human error with reduced data necessities (e.g., requiring merely one expert prediction per instance) and employs constraint programming to globally minimise error costs while adhering to workload restrictions. In particular, DeCCaF incorporates a component which simultaneously models the behaviour of the human team and forecasts the likelihood that deferring to a certain expert would provide a correct decision.

4.2.6 Strengths and limitations of Learning to Defer. In contrast to algorithms that operate under oversight (see Section 3), Learning to Abstain Hybrid Systems are trained not to predict when their performance is weak. As a result, when using an L2R or L2D algorithm to make decisions, one can expect to receive two kinds of evidence: the machine's prediction concerning the action of deferral and, if the AI does not abstain, the result of the prediction task. L2D algorithms improve upon L2R by incorporating a representation of human knowledge directly in the training process. In such a way, the deferral policy is trained to adapt to both the AI model and the human decision-maker, ideally the same that will employ the hybrid system. Recent empirical investigations involving human subjects yielded evidence for the additional advantages that abstaining systems bring to hybrid systems. Hemmer et al. [69] found that such algorithms improve both *human task performance* compared to a human or an AI working alone, and *human task satisfaction* compared to a human working alone. A different study [124] also investigated the effects of employing abstaining hybrid systems on the human perception of AI performance and credibility. The results indicate that users are frequently influenced by the system's recommendation also on ambiguous instances, even without conscious awareness, and thus support the adoption of L2D algorithms. However, L2D also comes with several limitations [94]. Most importantly, these include: data availability issues, which are primarily due to the need of human predictions in addition to the ground truth for all instances within the training set and all human agents involved in the Hybrid System; and fairness concerns that may stem from the introduction of bias by both human and machine agents, as well as from the abstention mechanism itself [76]. Although there have been suggestions to deal with such issues, these proposals still do not offer straightforward solutions.

In terms of the human's role in the Hybrid Decision Making System, Learning to Abstain only partially enhances the paradigm of human oversight over machines (Section 3), as the deferral remains exclusively a machine-side operation and there is no direct human-side interaction considered in the design of the algorithms.

5 Learning Together

The next natural step in hybrid systems is a two-way collaboration in which human agents are not mere executors or overseers but can directly *interact* with the machine. In Paradigm 1 (Section 3) we have highlighted the problems of oversight for automated systems. Abstention (Paradigm 2, Section 4) aims to mitigate these limitations by enabling the machine to abstain from unreliable prediction while, by incorporating a representation of human capabilities in the learning process, facilitating a fine-tuned selection of cases requiring human intervention. In Paradigm 3, or *Learning Together*, (LT) we present a discussion on the subsequent advancement in the development of synergistic HSs: rather than merely routing hard cases to humans, LT explicitly *integrates* human knowledge, reasoning, and feedback into the machine learning process, while having the machine unveil its own decision process. Having both reveal their decision-making process allows both of them to integrate the other's feedback, thus creating a looping system where the weaknesses of one are compensated by the strengths of the other. Thus, LT systems are designed to support *bidirectional* communication between human and machine and bidirectional integration of the other's expertise. This is in stark contrast with alignment models, e.g., Reinforcement Learning from Human Feedback (RLHF), where communication is one-directional, the decision process is opaque, no loop exists, and machines are merely trained to one-directionally replicate preferences provided by humans, rather than their decision-making process.

5.1 Towards humans taking control of the decision-making process

Two primary configurations of human–AI interaction can be distinguished based on the extent of human control in the decision-making process [119]: in one, the machine component is in control, utilising human inputs primarily to guide the model towards a better (potentially local) optimum; in the other, the human occupies the

central role and maintains full control, with the machine component merely providing suggestions or intermediate results within a decision-making framework for which the human retains full accountability.

Machines in control: interaction for machine learning. Architectures belonging to the first family of LT systems typically involve human participation primarily during the training phase, resting on the assumption that intelligent systems can generalize more effectively and handle previously unseen situations when they learn in close interaction with humans . In certain cases (e.g., in interactive machine learning), the learning process through human-AI interaction extends beyond training into the deployment stage, allowing for continuous and incremental model improvement that adapts to changing contexts. Note, however, that this ongoing adaptation is still aimed at optimizing the machine performance and cannot be considered part of the decision-making process per se [113], as it happens in the architectures described in Section 5.1. The defining distinction among the various learning paradigms within this family lies in the degree of control exercised by each agent over the learning process. In *Active Learning* [147], the system retains primary control and queries human teachers in the role of "oracles" to obtain labels for data instances that are uncertain or unknown to the model. Similarly, in Reinforcement Learning with Human Feedback [28] humans evaluate or reward the system's outputs during training, thereby guiding policy optimisation. By contrast, *Interactive Machine Learning* [138] reflects a more balanced form of collaboration, in which control is shared: humans can adopt various roles (e.g., domain expert, data scientist, or crowdworker) and engage at various stages of the workflow (e.g., initially in identification or annotation tasks or in the end to validate or correct the machine's output). At the opposite end of the spectrum, *Machine Teaching* [185] places control largely in the hands of human experts, who act as deliberate instructors, transferring their domain knowledge and conceptual understanding to the model with the explicit aim of steering its learning trajectory.

Humans in control: interaction for learning together. In Learning Together systems, the interaction is not designed merely to optimise machine performance but to enhance the decision-making process as a whole. While human agents can still participate in the learning process of the machine, they serve as primary decision-makers inside the system. This paradigm aligns with the principles of Human-Centred AI (HAI), which emphasises the integration of human conditions, contexts, and values into the design and evaluation of AI systems. Although these principles are fundamental to all hybrid systems, they achieve their closest realisation within the Learning Together paradigm: in this context, responsible design is inherently integrated into the system's learning dynamics, influencing the training, interaction, adaptation, and coevolution of human and machine agents over time. In practice, the design of an LT system begins with the development of trustworthy AI architectures that explicitly consider the broader social, ethical and cultural dimensions within which the system operates. Moreover, it needs the creation of effective user interfaces that promote usability, transparency and user engagement, ensuring that human participants can interact with the system intuitively and confidently. Finally, the evaluation of LT systems extends beyond traditional technical metrics such as accuracy or F1-score, incorporating human-centred criteria (such as interpretability, usefulness, fairness, and trustworthiness) that better reflect the system's real-world impact and its ability to support meaningful, equitable, and sustainable human decision-making [113]. Conceptually, the LT paradigm closes the feedback loop between human and machine. Through corrective actions, structured explanations, or partial labels, LT systems not only identify when humans are better decision-makers but also learn human reasoning patterns to improve future autonomy.

Scope of this section. To keep the scope of this section self-contained, this paradigm is explicitly *centred on discriminative prediction tasks (classification/structured prediction)* where (i) decisions are discrete or score-based and (ii) operational performance, calibration, and cost metrics are well defined. We therefore exclude (for the core LT narrative) most work that treats large language models (LLMs) as primary decision makers: while LLMs provide powerful communication primitives, an expanding empirical literature documents their hallucinations, calibration

failures, and brittleness in high-stakes decision support, problems that make them unreliable foundations for the sort of tightly coupled learning with humans we target here [47, 80]. LLM-centric methods and how they may be used safely are discussed separately in Section 6.

5.2 Algorithms for learning together

Learning Together (LT) creates a bidirectional channel in which human and machine explain their predictions to one another and learn from those explanations. This mutual interaction is aimed at strengthening the machine in terms of accuracy, generalisation, a shallower learning curve and more transparency, while it also benefits the human by providing a deeper understanding of how the machine solves the task instead of merely replacing them.

Example of melanoma detection: Learning Together (LT)

In Learning Together systems, the dermatologist is provided with machine-generated hypotheses enriched by interpretable explanations (such as highlighted visual regions, feature attributions or references to clinically meaningful patterns) and can actively interact with them by validating, correcting or refining the model's reasoning based on clinical expertise. Crucially, this interaction serves a dual purpose: it supports the clinician with transparent, data-driven evidence while simultaneously enabling the AI system to progressively adjust to the clinician's diagnostic reasoning. Over time, this bidirectional exchange results in a co-evolving hybrid system in which decision quality and trust are continuously strengthened.

Formally, the simplest LT model for a single human expert with side information Z can be defined as: $\hat{y} = f_\theta(X, Z)$ where X are machine-observed features, Z is (structured or unstructured) information produced by the human, and θ are model parameters updated through interactions[2]. There are many different ways in which an LT system can be implemented. The operational framework of LT systems can be characterised according to three properties: the *language*, namely, the structured system of communication that governs the bidirectional exchange of information between agents; the *time of interaction*, which defines when the exchange occurs; and *learning cost*, which quantifies the computational resources required for effective collaboration. Here we briefly discuss the different families of techniques that exemplify these properties and that can be used to build an HDMS.

5.2.1 Hard Reasoning Languages. Hard reasoning languages, i.e., logic languages, are natively both understandable to humans and relatively easy to use due to their similarity to human reasoning. Still, it is not straightforward to embed them in the machine learning model due to their symbolic nature, which is in stark contrast with the subsymbolic nature of most machines, e.g., neural models. A solution to this problem are *neurosymbolic* models with a subsymbolic and symbolic component [44, 63]. The subsymbolic component is typically a neural one, e.g., a neural network, which fully embodies machine reasoning. The symbolic component is geared towards the human and thus employs a logic language.

Logic languages encode knowledge as clauses/rules and facts; logic engines produce logical derivations as extrinsic artefacts that humans can act on (e.g., by adding or deleting rules, facts, or their compositions). Hard logic has strong theoretical guarantees, so it suits highly compliant systems where human feedback can be reliably integrated. Because logic is expert-driven rather than data-driven, artefacts are often hand-crafted, domain-specific, and costly to automate.

Figure 3 shows a CLEVR example [75]: a program answers the question *"What is the colour of the cube to the right of the yellow sphere?"* by locating the yellow sphere, selecting objects to its right, filtering by shape, and

[2]For the sake of simplicity, here we model a single human agent, but the formulation can be extended to multiple agents.

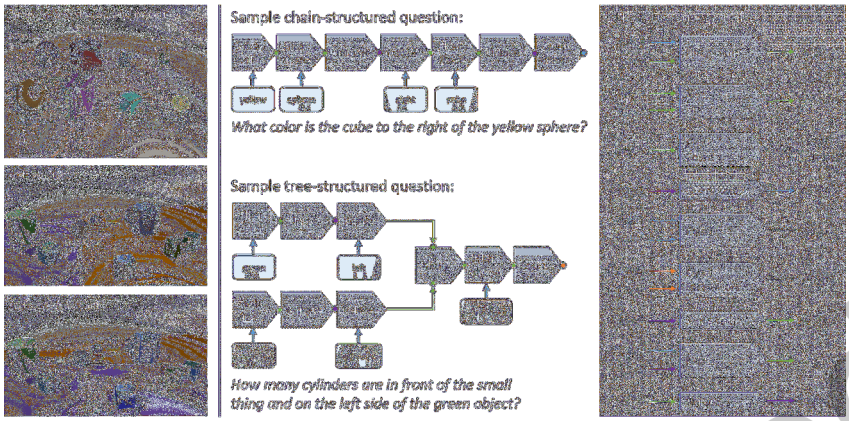


Fig. 3. An example of Question Answering machine with a hard reasoning language. The agent maps the question to a program using a set of primitives and then executes the program on the input, providing the human with both a prediction and a malleable program that they can correct.

extracting the colour, that is, an inspectable/modifiable artefact. A similar hard-reasoning instance appears in [63], where a machine predicts a Tic-Tac-Toe outcome using logic rules that humans may inspect and edit.

5.2.2 Soft reasoning languages. Soft reasoning languages improve the flexibility of hard reasoning languages at the cost of their strong theoretical properties by replacing logic rules and facts with *logic-like* rules and the symbolic reasoning engine with a subsymbolic approximate one. These languages are almost exclusively Natural Languages, and the machines leveraging them are language models.

They originate from the work by Kassner et al. [78], and build on top of two other families, knowledge injection models [95] and soft reasoners [31], the former providing models able to integrate external, possibly human-sourced, knowledge in their inference, and the latter providing soft reasoning engines for Natural Language.

Injections can be in the form of simple logic-like rules [112], or full reasoning trees [36]. Like in hard reasoning languages, possible actions include *addition* and *deletion* of new/existing rules or facts to the model. In Large Language Models, prompt engineering can be used at inference time to interact with the model and solve a given task. Chain models [175] are a perfect example of such a case, as they allow a plethora of different interactions. In Chain of Thoughts [171], the machine generates a straightforward sequence of instructions, named "thoughts", on how to solve the task, i.e., some reasoning steps or derivations. In Chain of Command [180], the machine may task artefact generation to other machines, while in Chain of Question [72] artefacts are retrieved from external knowledge bases. In Tree of Thoughts [178], multiple possible derivations are considered simultaneously, thus simulating thought backtracking, beam search, and multiple alternative solutions to the task. In Chain of Instructions [66], previous steps are reused in future derivations. A key advantage of soft reasoning languages is their ability to combine artefacts from many sources, and Knowledge Graphs (KGs) are a primary example. KGs are widely available, domain-spanning, can be mined from and verbalised into natural language, are human-understandable and often community-verified (so require little extra checking and scale to many users), and integrating them into systems is a well-studied problem with effective solutions [95]. Knowldege Graphs also allow some basic form of reasoning which can be easily extended: they provide strict reasoning, both commonsense [74] and factual [165], all of which can be integrated in Natural Language.

Figure 4 presents an example from [167]. Here, the goal is to understand what the purpose of the red object (the hydrant) in the photo is. We ought to remark on the importance of the context, since even though the

Question: What can the red object on the ground be used for?
Answer: Firefighting.
Support fact: Fire hydrants can be used to fight fires.

Fig. 4. An example of a Question Answering hybrid system employing a soft reasoning language and leveraging Knowledge Graphs [167]. Here, the machine consults its artefact bank to retrieve a supporting fact for its prediction and provides it to the human agent alongside its prediction.

machine is aware of the concept of "fire hydrant", it is highly unlikely to have seen either a fire hydrant or a fireman in a forest. Moreover, since the image itself does not provide any information regarding the purpose of the hydrant, one should expect that the machine could not solve the task without the additional information provided by the human agent in the Knowledge Graph. Retrieved facts can also be leveraged in Chain models for Retrieval-Augmented Generation (RAG), with systems employing retrieved facts for Chain of RAGs [184], Tree of RAGs [84], or Chain of Search [176].

5.2.3 Explanation Languages. Explanations are designed to *explain* the machine, which makes them highly understandable artefacts out of the box. Common families of explanations include counterfactuals, prototypes, feature relevance, and decision rules, the last two leveraged in LT systems. Feature relevance provides the human agent with the estimated influence that each input feature has on the prediction of the machine: the higher the relevance, the higher the sensitivity of the machine to changes in that feature. Decision rules, instead, provide a descriptive understanding by articulating with logic rules the predictions of the machine. In all the following approaches, the key point is that explanations are used to trigger an internal change by the machine: the human interacts through the explanations, and the machine modifies the model based on the explanations.

Of particular interest are post-hoc, model-agnostic explanations, which are extracted *after* the machine has been trained, and regardless of its form. Jointly, these two characteristics minimise coupling between the interaction artefacts and the machine, thus granting more flexibility in the design of the machine. On this account, explanations are major artefacts of interest, particularly for systems in which the interaction is the focal point, such as Interactive Machine Learning [4], and eXplainable Interactive Learning (XIL) [159] systems. In particular, the latter allows the human agent to inspect and provide feedback to the machine by correcting its explanation. An XIL machine is based on a simple algorithmic kernel comprised of five steps:

(1) the machine performs a *learning* step by optimizing its parametrization θ;
(2) the machine generates explanations e_M of (a subset of) its predictions;
(3) the human examines e_M, and provides a (optionally) corrected explanation e_H;
(4) the machine performs a learning step by optimizing θ according to the corrected explanation;
(5) if no stopping criterion is met, the machine returns to step (1).

By iteratively querying the human, the machine parametrisation is thus conditioned on the corrected explanations e_H, that is, an XIL machine ends up implementing an estimator of the form: $f_{\theta|e_H}(X)$, where the correction e_H acts as a proxy for the human knowledge Z.

Integration is directly dependent on the family of explanations. By far the most widespread one is feature importance, which assigns a relevance score to each element of the input data X.. The interaction then consists

of a possible correction of the relevance scores, with the human agent activating or deactivating each feature according to their judgement. Integration follows either a *learning* approach, in which the correction is directly encoded in the machine training objective [140], or a *generative* one, where the correction is implemented via training on additional synthetic data [158, 159]. In a learning approach, the corrections of the human agent are encoded in a correction matrix $C \in \{0,1\}^{n \times m}$ that states what feature relevance has been corrected for each single instance. In a generative approach, C is instead used to generate synthetic data $\widetilde{X}$ to further train the machine. Features with low relevance have randomised or copied values, while features with high relevance are kept as-is [159]. A small subset of machines are designed to explicitly encode feature relevance in their architecture, thus allowing direct manipulation by human agents [169].We may also have a generative approach in which the human agent is tasked with directly creating the additional synthetic data themselves, working on global explanations [129].

In all the aforementioned approaches, the key point is that explanations are used to trigger a new learning phase for the machine: the human interacts through the explanations, and the machine retrains based on the explanations. A similar approach to interaction through explanations can be employed with systems that are built to be interpretable by design.

For example, in [87] we have a learning approach in which the explanation is itself a component of the architecture; thus, the correction is an added component of the machine objective. An emerging approach, mainly aimed at *Concept Explanations*, is the *structured explanation* approach, where explanations are provided as complex structures that the human can act upon. [17] presents an application on concept hierarchies, where concepts are laid on a tree-like hierarchy such that the concept of a parent node, e.g., "Animal", is a generalisation of the concepts in its children, e.g., "Dog" and "Cat". The machine, based on a k-NN model, is tasked to solve two tasks: a downstream task and a *concept drift* task, that is, to identify if the relationships within the structure have changed. Once detected, the machine presents the concepts of interest to the human, who in turn corrects their structure, e.g., by removing or adding concepts, or by acting on the structure itself, i.e., removing or adding parent-child relationships between concepts. The correction is integrated by removal/addition of appropriate instances from the training set of the machine, thus directly impacting the k-NN model.

5.2.4 Time of interaction and Cost Of Learning. Artefacts are integrated at different times in the lifetime of the machine, hence either at training or at inference time. In the former case, the machine integrates the artefact at training time and cannot be interacted with at inference time. In this case, the human agent is effectively providing feedback only on training [17, 87, 129, 140, 158, 159, 169]. In the latter case, the human agent has more control over the machine and receives and acts upon interaction artefacts at inference time [18, 44, 63, 112, 116, 167]. Machines based on explanation languages, such as XIL, tend to provide training-time interaction, while more recent approaches based on hard or soft reasoning languages tend to provide inference-time interaction.

A critical distinction in LT systems is the cost of learning, that is, the cost the system has to pay to properly integrate the actions of the human within the machine. Systems tend to fall to the two ends of the spectrum. In traditional approaches such as XIL, the interaction triggers a costly training step for the machine. Here, the cost varies with respect to both the magnitude of the human agent correction and the intrinsic features of the machine [17, 87, 129, 140, 158, 159, 169]. Conversely, more recent approaches [18, 63, 112, 116, 167], such as extrinsic artefact-based systems, have no additional cost due to the nature of the machine itself. Why then rely on traditional approaches if they incur an inevitable additional cost? Extrinsic artefacts have to be generated in the first place: the cost of populating the artefact bank is as inevitable as the training cost for traditional approaches. As previously mentioned, when such banks already exist, e.g., in knowledge graphs, this cost can be greatly reduced.

5.2.5 Strengths and limitations of Learning Together Systems. Unlike Human overseers and Learning to Abstain, Learning Together systems integrate humans and machines in a hybrid, bidirectional system. The design of agents, tasks, domain, data are often coupled, with one notable exception in soft reasoning systems, where convergence is occurring. Machines are often limited by the chosen language, meaning *i)* system design requires significant effort, and *ii)* communication does not easily transfer across languages. This heterogeneity hinders progress, with each system offering insights mainly for similar tasks, domains, and interactions. Thus, Learning Together systems improve *vertically*, with unpredictable success.

Currently, Learning Together systems are largely static: they cannot switch languages on demand, adapt to different human agents, or defend against incorrect human feedback. As a recent development in hybrid systems, they also lack tailored validation measures. While traditional validation applies to the system as a whole, little effort has been made to assess how well a machine complies with human corrections or to provide theoretical guarantees on the effects of those corrections.

6 LLMs for Hybrid Decision Making

Large Language Models (LLMs) are a modern development of artificial intelligence that allow humans to interact with powerful generative models using natural language. This empowers even non-expert users with the ability to elicit desired answers from such models. While a full survey on how LLMs are used in human-AI collaboration is outside the scope of our work, we want to contextualise how LLMs can be employed in hybrid decision-making systems. LLMs are compelling because they produce fluent, contextually rich language and integrate knowledge from disparate sources, producing outputs that appear highly coherent and plausible. Yet those same properties make them a poor fit as-is for many decision pipelines. Empirically, LLMs can generate factually incorrect yet persuasive statements ("hallucinations") and often verbalise their outputs with undue confidence; both phenomena undermine reliability in domains where a single wrong suggestion can be consequential (e.g., medicine, law). Clinical evaluations demonstrate that providing physicians with LLM assistance does not automatically improve diagnostic reasoning and, in some settings, fluent but incorrect prose can mislead clinicians unless interfaces surface uncertainty and provenance clearly [56, 149, 187]. From a systems and human factors perspective, LLMs are also opaque: for closed-source commercial models, the user typically cannot inspect internal representations or access reliable numeric confidence scores, making standard safety measures (e.g., calibrated uncertainty, provenance tracing) harder to implement. LLMs are sensitive to prompt formulation and distributional shifts; small changes in phrasing or data distribution can produce large differences in output quality, increasing brittleness in real workflows. Moreover, because most deployed LLMs are updated via offline, curated procedures (supervised fine-tuning, RLHF), corrections made during an interactive session do not normally change the model's internal parameters immediately, which complicates the "teach while you use" expectation many practitioners have for human-in-the-loop systems [123]. Finally, overreliance on persuasive text can erode human skills (deskilling) or encourage automation complacency unless the workflow deliberately preserves human verification [56, 100]. Regarding our taxonomy of hybrid decision-making systems, we find works and applications of LLMs in each paradigm.

Human Oversight. In the human-oversight configuration humans review AI outputs and make the final decision. Several controlled clinical studies exemplify this arrangement: randomised trials and observational work that inserted a conversational LLM into physicians' toolset but left decision authority with the clinician. For example, Goh et al. evaluated a commercial LLM (GPT-4 via ChatGPT Plus) as an aid for physicians and found no significant improvement in diagnostic reasoning over conventional resources; the study highlights both potential value and the crucial role of clinician verification and contextual judgement when LLMs are used as assistants rather than decision makers [56]. Likewise, empirical assessments of LLM performance on clinical text extraction show

hallucination and inconsistency risks that must be mitigated by human review and careful UI affordances [149]. Taken together, these findings suggest that human supervision remains essential even in LLM-based systems and, moreover, that the inclusion of LLM components may introduce additional challenges for oversight, calling for increased attention to interface designs that clearly convey model uncertainty and limitations while ensuring the underlying reasoning process is transparent and interpretable.

Learning to Abstain. The literature on selective classification and abstention for LLMs is rapidly developing [173]. Experimental work shows some promising concrete methods: self-evaluation token-level strategies improve selective generation performance [139]; conformal and PAC-style selection functions enable controllable abstention guarantees in selective generation [92]; inference-time strategies also exist to reduce over-abstention in vision-language systems by using LLMs to seek additional evidence before giving up [153]. These studies indicate that abstention is a promising solution that is able to reduce high-consequence failures, but (i) calibration is nontrivial because many LLMs are overconfident, and (ii) abstention policies must be carefully designed to avoid excessive conservatism or biased routing of difficult cases [173, 181].

Learning Together and Alignment. Humans can shape LLM behaviour through several complementary *alignment* mechanisms that operate at different timescales and with different guarantees , some operating at training, others at inference time. Training-focused mechanisms such as fine-tuning, reinforcement learning from human feedback (RLHF), and direct preference optimisation (DPO), collect datasets of human behaviour, expertise, and preferences, i.e., labels, explanations, and preference rankings, and align the machine to the human through a training step on such dataset. These methods are the most prevalent in the alignment of large deployed assistants to user intent, as they enable centralised quality control and auditability. Inference-focused alignment mechanisms include few-shot prompting (FSP) via in-context learning [42], wherein the human expertise is provided via examples or instructions at inference time, and the machine exploits the knowledge gathered in pretraining . This is fast and flexible but brittle and often sensitive to exemplar choice, order, and phrasing , and it does not produce *lasting* model changes. A related approach is chain-of-thought prompting (COT) [29] (cfr. Section 5), wherein the machine also unveils its own decision process. Still, none of the aforementioned mechanisms fit the Learning Together system definition provided in our taxonomy due to their lack of bidirectional communication (DPO, FSP), unveiling of each agent's reasoning (RLHF, FSP), or ephemeral effects of the interaction (FPS, COT). That is, the distinction between Learning Together and alignment mechanisms is rooted in deep structural differences between alignment methods and LT systems. In Learning Together, the human expertise is part of an interactive closed-loop decision-making process in which human actions directly condition the system's behaviour, *and viceversa*. By contrast, alignment methods merely aim to align the machine's behaviour to the human's and have the machine's decision process replace the human's entirely. At their core, alignment and Learning together differ on their goal: the former aims to have machines mimicking humans, while the latter aims to have cohesive and complete systems integrating both.

Example of AI-assisted programming: LLMs

In AI-assisted programming environments, the three hybrid paradigms naturally coexist within a single workflow. Under Human Oversight, the language model proposes code snippets or explanations that the developer reviews, edits, or discards, retaining full decision authority. Learning to Abstain emerges when the model signals uncertainty, requests clarification, or refrains from producing code for underspecified or ambiguous tasks, deferring control back to the human. Learning Together arises when the developer and the model interact, e.g., through refinement of partial solutions, constraints, and corrections provided by the human, that condition subsequent model outputs, leading to a co-evolving decision process.

LLMs in high-stakes decision-making. Deploying state-of-the-art proprietary LLMs in domains regulated by GDPR, HIPAA, or equivalent rules raises multiple, practically significant obstacles. First, using remote, closed-box APIs for sensitive patient or client data can run afoul of data-transfer, controller/processor, and data minimisation principles unless contractual, technical, and organizational safeguards are in place; regulators and expert guidance documents explicitly call out LLMs as requiring tailored risk assessments and mitigation measures [46]. Second, proprietary models typically run as hosted services (inference via API), which complicates the required data flow documentation and may expose personal data to third parties or cross-border transfer rules. Recent peer-reviewed guidance and reviews for health data advise local/private hosting, differential privacy, synthetic data generation, split-inference architectures, or cryptographic protocols (e.g., MPC, secure enclaves) as pragmatic mitigations for EHR and other sensitive data uses [77, 100]. However, these mitigations trade off capability: smaller locally hosted models or heavily DP-protected models often deliver lower performance than the top proprietary systems, and the cost/engineering complexity of private deployment can be significant. Consequently, strict privacy regimes frequently push practitioners toward simpler models or hybrid architectures to remain compliant while balancing utility and risk [46, 77].

7 Literature gaps and concluding remarks

Hybrid Decision-Making Systems, where humans and machines collaborate on predictive tasks, mark a new AI paradigm. Synergistically integrating different agents allows us to leverage their strengths and mitigates their respective weaknesses. This survey introduces a taxonomy of key Hybrid Systems literature, categorised into three paradigms: Human Oversight, Learning to Abstain, and Learning Together. Human Oversight relies on passive human supervision of autonomous machine decisions; Learning to Abstain introduces a routing mechanism that dynamically allocates decisions between human and machine; lastly, Learning Together represents the deepest integration, where human expertise is directly incorporated into the learning process through sustained interaction, and vice versa, enabling mutual adaptation over time. We reviewed representative works within each paradigm, discussing their respective strengths and limitations in order to inform and support future research on HDMS. Building on this taxonomy, we identify four broad categories of open challenges.

Machine-related challenges. Unlike Learning Together systems, Human Oversight and Learning to Abstain systems still lack meaningful means of communication to engage with the human agent. When presented with a prediction to oversee, or with an uncertainty estimate of such a prediction, human agents are rarely also presented with suggestions on why the prediction should be accepted or rejected, or why the machine is uncertain of its prediction, let alone how to tackle the uncertainty itself. We have highlighted some first steps in tackling this problem in Subsection 3.2, but this is far from a solved problem.

Human-related challenges. Both the Learning to Defer and Learning Together systems require a considerable human effort, i.e., a high volume of labels or artefacts, to be implemented. This poses a significant burden on the human agent, and while current solutions aim to tackle the problem with a given fixed budget, there are no clear solutions as to how to reduce such high costs. When it comes to Learning to Defer, human-AI "collaboration" heavily depends on a global data annotation industry, wherein a vast and often invisible human labour force is engaged with tedious and taxing jobs. Yet, the prevailing labour standards for data annotators are mostly characterised by lax regulations, low wages and few legal protections [143].

Interaction-related challenges. With respect to interaction, there is limited understanding of which communication languages are best suited to different tasks and how they should adapt to diverse users within the same hybrid system. Most existing systems are effectively *monolingual*, designed for a single type of human agent, and implicitly assume static human capabilities and understanding. This assumption limits flexibility and prevents adaptation to heterogeneous users. These limitations are exacerbated in settings involving multiple human agents,

such as Multiple-Expert Learning to Defer and Learning Together systems, where challenges related to agent multiplicity, conflicting inputs, and malicious or adversarial behaviour remain largely unexplored.

Human-AI coevolution challenges. Hybrid Systems are embedded in broader sociotechnical environments in which humans and AI coevolve over time within societal structures. At the micro level, repeated interactions with AI shape human behavior, skills, trust, and decision strategies, while at the macro level the deployment of hybrid systems influences, and is influenced by, economic, legal, political, and regulatory contexts. These feedback loops introduce long-term dynamics that are largely overlooked in current hybrid system designs [126]. From a technical perspective, a key open challenge lies in the development of AI architectures capable of adapting to evolving human factors over time, including changes in goals, intentions, expertise, learning effects, and physical or contextual conditions [183]. Beyond technical considerations, hybrid systems face significant non-technical challenges. Legal frameworks impose requirements on transparency and responsibility, governance structures raise questions about oversight and societal impact, and socio-economic concerns include labour effects and unequal access to AI technologies. A key open problem is how to translate these regulatory and societal requirements into concrete, operational mechanisms within hybrid systems. Although technical solutions may exist in principle, the rapid adoption of language-model-based systems raises doubts about whether current approaches can reliably satisfy these constraints.

Acknowledgments

This work has been supported by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. This work was also funded by the European Union under ERC-2018-ADG Grant Agreement no. 834756 "XAI: Science and technology for the eXplanation of AI decision making" and under Grant Agreement no. 101120763 - TANGO. 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

[1] Alex Albright. 2019. If you give a judge a risk score: evidence from Kentucky bail decisions. *Law, Economics, and Business Fellows' Discussion Paper Series* 85 (2019).

[2] Saar Alon-Barkat and Madalina Busuioc. 2023. Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory* 33, 1 (2023).

[3] Jean V. Alves, Diogo Leitão, Sérgio M. Jesus, Marco O. P. Sampaio, Javier Liébana, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. 2024. Cost-Sensitive Learning to Defer to Multiple Experts with Workload Constraints. *Trans. Mach. Learn. Res.* (2024).

[4] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Mag.* 35, 4 (2014).

[5] Shin Ando and Ayaka Yamamoto. 2023. Anomaly Detection via Few-Shot Learning on Normality. In *Machine Learning and Knowledge Discovery in Databases*, Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas (Eds.). Springer International Publishing, Cham, 275–290.

[6] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2022. Conformal Risk Control. (2022).

[7] Amina Asif and Fayyaz ul Amir Afsar Minhas. 2020. Generalized Neural Framework for Learning with Rejection. In *2020 International Joint Conference on Neural Networks (IJCNN)*.

[8] Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2022. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*. PMLR, 1117–1174.

[9] Alexander Babuta and Marion Oswald. 2021. Machine learning predictive algorithms and the policing of future crimes: Governance and oversight. In *Predictive Policing and Artificial Intelligence*.

[10] Lisanne Bainbridge. 1983. Ironies of automation. *Autom.* 19, 6 (1983).

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *AAAI Conference on Artificial Intelligence*.

[12] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *HCOMP*.
[13] Gagan Bansal and Daniel S. Weld. 2018. A Coverage-Based Utility Model for Identifying Unknown Unknowns. In *AAAI*.
[14] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. 2006. Convexity, Classification, and Risk Bounds. *J. Amer. Statist. Assoc.* (2006).
[15] Peter L. Bartlett and Marten H. Wegkamp. 2008. Classification with a Reject Option Using a Hinge Loss. *JMLR* (2008).
[16] Roberto Battiti and Anna Colla. 1994. Democracy in neural nets: Voting schemes for classification. *Neural Networks* (1994).
[17] Andrea Bontempelli, Fausto Giunchiglia, Andrea Passerini, and Stefano Teso. 2022. Human-in-the-loop handling of knowledge drift. *Data Min. Knowl. Discov.* 36, 5 (2022).
[18] Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, and Greg Durrett. 2021. Flexible Generation of Natural Language Deductions. In *EMNLP*.
[19] Johannes Brinkrolf and Barbara Hammer. 2017. Probabilistic extension and reject options for pairwise LVQ. In *International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization*.
[20] Johannes Brinkrolf and Barbara Hammer. 2018. Interpretable machine learning with reject option. *at - Automatisierungstechnik* (2018).
[21] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (2021).
[22] Federico Cabitza and Chiara Natali. 2022. *Open, Multiple, Adjunct. Decision Support at the Time of Relational AI.* IOS Press. doi:10.3233/faia220204
[23] Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. 2024. In defense of softmax parametrization for calibrated and consistent learning to defer. *Advances in Neural Information Processing Systems* 36 (2024).
[24] Hubert Cecotti and Szilárd Vajda. 2013. Rejection Schemes in Multi-class Classification - Application to Handwritten Character Recognition. In *ICDAR*. IEEE Computer Society, 445–449.
[25] Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. 2022. Sample Efficient Learning of Predictors that Complement Humans. In *Proc. of the 39th International Conference on Machine Learning*.
[26] Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023. Say What You Mean! Large Language Models Speak Too Positively about Negative Commonsense Knowledge. In *ACL (1)*.
[27] C. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16, 1 (1970).
[28] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.
[29] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. In *Annual Meeting of the Association for Computational Linguistics*.
[30] Manuel R. Ciosici, Joe Cecil, Dong-Ho Lee, Alex Hedges, Marjorie Freedman, and Ralph M. Weischedel. 2021. Perhaps PTLMs Should Go to School - A Task to Assess Open Book and Closed Book QA. In *EMNLP (1)*.
[31] Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as Soft Reasoners over Language. In *IJCAI*.
[32] Lize Coenen, Ahmed K. A. Abdullah, and Tias Guns. 2020. Probability of default estimation, with a reject option. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (2020).
[33] David Cohn, Les Atlas, and Richard Ladner. 2016. Improving generalization with active learning. *Machine Learning* (2016).
[34] Stuart Coles. 2001. *An introduction to statistical modeling of extreme values*. London.
[35] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. 2016. Learning with rejection. In *International Conference on Algorithmic Learning Theory*. Springer.
[36] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining Answers with Entailment Trees. In *EMNLP (1)*.
[37] A. P. Dawid. [n. d.]. The Well-Calibrated Bayesian. *J. Amer. Statist. Assoc.* ([n. d.]).
[38] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. 2020. Regression Under Human Assistance. In *AAAI Conference on Artificial Intelligence*.
[39] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez-Rodriguez. 2020. Classification Under Human Assistance. In *AAAI Conference on Artificial Intelligence*.
[40] C. De Stefano, C. Sansone, and M. Vento. 2000. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 30, 1 (2000).
[41] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015).
[42] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1107–1128. doi:10.18653/v1/2024.emnlp-main.64

[43] Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-free Selective Classification. *JMLR* (05 2010).
[44] Kevin Ellis, Catherine Wong, Maxwell I. Nye, Mathias Sablé-Meyer, Lucas Morales, Luke B. Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B. Tenenbaum. 2021. DreamCoder: bootstrapping inductive program synthesis with wake-sleep library learning. In *PLDI*.
[45] Birte Englich, Thomas Mussweiler, and Fritz Strack. 2006. Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin* 32, 2 (2006).
[46] European Data Protection Board (EDPB). 2025. *AI Privacy Risks & Mitigations — Large Language Models (LLMs).* Technical Report. https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf
[47] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.* 630, 8017 (2024), 625–630.
[48] Andrew Guthrie Ferguson. 2020. High-tech surveillance amplifies police bias and overreach. *The Conversation* 12 (2020).
[49] Lydia Fischer, Barbara Hammer, and Heiko Wersing. 2014. Local rejection strategies for learning vector quantization. In *ICANN '14*. Springer.
[50] L. Fischer, Barbara Hammer, and H. Wersing. 2015. Efficient rejection strategies for prototype-based classification. *Neurocomputing* 169 (04 2015).
[51] Vojtech Franc and Daniel Prusa. 2019. On discriminative learning of prediction uncertainty. In *ICML*.
[52] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI Collaboration with Bandit Feedback. In *International Joint Conference on Artificial Intelligence*.
[53] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 30.
[54] Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *International Conference on Machine Learning*.
[55] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.*
[56] Ellen Goh et al. 2024. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open* 7, 10 (2024), e2440969. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395
[57] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. http://www.deeplearningbook.org.
[58] Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. 2008. Support Vector Machines with a Reject Option. In *Advances in Neural Information Processing Systems*, Vol. 21.
[59] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proc. of the 2018 WWW*.
[60] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. 2022. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. In *Equity and Access in Algorithms, Mechanisms, and Optimization*.
[61] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intell. Syst.* 34, 6 (2019).
[62] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (2019).
[63] Lijie Guo, Elizabeth M. Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart P. Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *IUI 2022*.
[64] László Györfi, Z. Györfi, and István Vajda. 1979. Bayesian decision with rejection. *Problems of Control and Information Theory* (1979).
[65] Lars Kai Hansen, Christian Liisberg, and Peter Salamon. 1997. The Error-Reject Tradeoff. *Open Systems & Information Dynamics* 4, 2 (April 1997).
[66] Shirley Anugrah Hayati, Taehee Jung, Tristan Bodding-Long, Sudipta Kar, Abhinav Sethy, Joo-Kyung Kim, and Dongyeop Kang. 2025. Chain-of-Instructions: Compositional Instruction Tuning on Large Language Models. In *AAAI*. AAAI Press, 24005–24013.
[67] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. 2022. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*.
[68] Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. 2023. Learning to Defer with Limited Expert Predictions. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 5 (6 2023).
[69] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 11 pages.
[70] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2024. Machine learning with a reject option: a survey. *Machine Learning* 113, 5 (2024).
[71] Victoria J. Hodge and Jim Austin. 2014. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review* (2014).
[72] Qiang Huang, Feng Huang, Dehao Tao, Yuetong Zhao, Bingkun Wang, and Yongfeng Huang. 2024. CoQ: AN Empirical Framework for Multi-hop Question Answering Empowered by Large Language Models. In *ICASSP*. IEEE, 11566–11570.

[73] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John C Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Position: TRUSTLLM: trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 813, 105 pages.

[74] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *AAAI*.

[75] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*.

[76] Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2021. Selective Classification Can Magnify Disparities Across Groups. In *ICLR*. OpenReview.net.

[77] J. Jonnagaddala et al. 2025. Privacy-preserving strategies for electronic health records in the era of large language models. *npj Digital Medicine* (2025). doi:10.1038/s41746-025-01429-0

[78] Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. In *EMNLP 21*.

[79] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* 6, 4 (2012).

[80] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2025. A Survey of Reinforcement Learning from Human Feedback. *Trans. Mach. Learn. Res.* 2025 (2025).

[81] Hendrik Kempt, Jan-Christoph Heilinger, and Saskia K Nagel. 2022. "I'm afraid I can't let you do that, Doctor": meaningful disagreements with AI in medical contexts. *AI & society* (2022).

[82] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. 2021. Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration. In *Advances in Neural Information Processing Systems*, Vol. 34.

[83] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2021. Towards Unbiased and Accurate Deferral to Multiple Experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

[84] M. Abdul Khaliq, P. Chang, M. Ma, Bernhard Pflugfelder, and F. Miletic. 2024. RAGAR, Your Falsehood RADAR: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models. *CoRR* abs/2404.12065 (2024).

[85] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[86] Riikka Koulu. 2020. Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law* 27, 6 (2020).

[87] I. Lage and F. Doshi-Velez. 2020. Learning Interpretable Concept-Based Models with Human Feedback. *Workshop on Human Interpretability in Machine Learning (ICML '20)* 1 (2020), 1–11.

[88] Thomas A. Lampert, André Stumpf, and Pierre Gançarski. 2016. An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. *IEEE Transactions on Image Processing* 25, 6 (2016).

[89] Thomas Landgrebe, David Tax, Pavel Paclik, and Robert Duin. 2006. The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters* 27 (06 2006).

[90] Hoel Le Capitaine and C. Frelicot. 2010. An Optimum Class-Rejective Decision Rule and Its Evaluation. *International Conference on Pattern Recognition* (08 2010).

[91] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004).

[92] M. Lee et al. 2024. Selective Generation for Controllable Language Models. In *Advances in Neural Information Processing Systems (NeurIPS) 2024*. https://proceedings.neurips.cc/paper/2024/file/5a6815122f533193a022cbc41786c1cc-Paper-Conference.pdf

[93] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018).

[94] Diogo Leitão, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. 2022. Human-AI Collaboration in Decision-Making: Beyond Learning to Defer. (2022).

[95] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS 2020*.

[96] Dongyun Lin, Lei Sun, Kar-Ann Toh, Jing Bo Zhang, and Zhiping Lin. 2018. Twin SVM with a reject option through ROC curve. *Journal of the Franklin Institute* 355, 4 (2018).

[97] Yi Lin. 2004. A note on margin-based loss functions in classification. *Statistics & Probability Letters* (June 2004).
[98] Jessie Liu, Blanca Gallego, and Sebastiano Barbieri. 2022. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. *Scientific Reports* 12, 1 (Feb. 2022).
[99] Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. 2024. Mitigating Underfitting in Learning to Defer with Consistent Losses. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4816–4824.
[100] Z. Liu et al. 2024. Surviving ChatGPT in healthcare. *npj Digital Medicine / PubMed Central (review)* (2024).
[101] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019).
[102] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS*.
[103] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, Vol. 31.
[104] Atefeh Mahdavi and Marco Carvalho. 2021. A Survey on Open Set Recognition. In *IEEE International Conference on Artificial Intelligence and Knowledge Engineering*.
[105] Frank Main. 2016. Cook County judges not following bail recommendations: study. *Chicago Sun-Times* (2016).
[106] Gianclaudio Malgieri and Giovanni Comandé. 2017. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law* 7, 4 (2017).
[107] Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. [n. d.]. Two-Stage Learning to Defer with Multiple Experts. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.).
[108] Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2024. *Principled Approaches for Learning to Defer with Multiple Experts*. Lecture Notes in Computer Science, Vol. 14494. Springer, 107–135.
[109] Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2024. Regression with Multi-Expert Deferral. *arXiv preprint arXiv:2403.19494* (2024).
[110] Natalia Markovich and Marijus Vaičiulis. 2023. Extreme Value Statistics for Evolving Random Networks. *Mathematics* (2023).
[111] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. Article 627, 10 pages.
[112] Bhavana Dalvi Mishra, Oyvind Tafjord, and Peter Clark. 2022. Towards Teachable Reasoning Systems: Using a Dynamic Memory of User Feedback for Continual System Improvement. In *EMNLP '22*.
[113] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2022. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (Aug. 2022), 3005–3054. doi:10.1007/s10462-022-10246-w
[114] Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. 2023. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. (2023).
[115] Hussein Mozannar and David Sontag. 2020. Consistent Estimators for Learning to Defer to an Expert. In *ICML 2020*.
[116] Shikhar Murty, Christopher D. Manning, Scott M. Lundberg, and Marco Tulio Ribeiro. 2022. Fixing Model Bugs with Natural Language Patches. In *EMNLP*.
[117] Yifat Nahmias and Maayan Perel. 2021. The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. *Harv. J. on Legis.* 58 (2021).
[118] Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. 2022. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 29292–29304.
[119] Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. 2025. Human-in-the-loop or AI-in-the-loop? automate or collaborate?. In *39th AAAI Conference on Artificial Intelligence and 37th Conference on Innovative Applications of Artificial Intelligence and 15th Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, Article 3185, 7 pages.
[120] Matey Neykov, Jun S. Liu, and Tianxi Cai. 2016. On the Characterization of a Class of Fisher-Consistent Loss Functions and its Application to Boosting. *Journal of Machine Learning Research* 17, 70 (2016).
[121] Navid Nobani, Fabio Mercorio, and Mario Mezzanzanica. 2021. Towards an Explainer-agnostic Conversational XAI. In *IJCAI*.
[122] Nastaran Okati, Abir De, and Manuel Rodriguez. 2021. Differentiable Learning Under Triage. In *Advances in Neural Information Processing Systems*, Vol. 34.
[123] Long Ouyang, Jeff Wu, Xue Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of NeurIPS 2022*. https://proceedings.neurips.cc/paper/2022/file/b1efde53be364a73914f58805a001731-Paper.pdf
[124] Andrea Papenmeier, Daniel Hienert, Yvonne Kammerer, Christin Seifert, and Dagmar Kern. 2023. Know What Not To Know: Users' Perception of Abstaining Classifiers. In *2023 ACM Designing Interactive Systems Conference (DIS 2023)*.
[125] Jagdish Parikh, Alden Lank, and Friedrich Neubauer. 1994. *Intuition: The new frontier of management*.
[126] Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, Alistair Knott, Yannis Ioannidis, Paul Lukowicz, Andrea Passarella, Alex Sandy Pentland, John Shawe-Taylor, and Alessandro Vespignani. 2025. Human-AI coevolution. *Artificial Intelligence* 339 (2025), 104244.

doi:10.1016/j.artint.2024.104244
[127] Emma Pierson. 2017. Gender differences in beliefs about algorithmic fairness. *CoRR* abs/1712.09124 (2017).
[128] Ignazio Pillai, Giorgio Fumera, and Fabio Roli. 2013. Multi-label classification with a reject option. *Pattern Recognition* (2013).
[129] Teodora Popordanoska, Mohit Kumar, and Stefano Teso. 2020. Machine Guides, Human Supervises: Interactive Learning with Global Explanations.
[130] M. Pradier, J. Zazo, S. Parbhoo, R. Perlis, M. Zazzi, and F. Doshi-Velez. 2021. Preferential Mixture-of-Experts: Interpretable Models that Rely on Human Expertise as Much as Possible. *AMIA Jt Summits Transl Sci Proc* 1 (2021).
[131] Andrea Pugnana and Salvatore Ruggieri. 2023. AUC-based Selective Classification. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, Vol. 206. PMLR.
[132] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*.
[133] Manish Raghavan, Solon Barocas, Jon M. Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *FAT* '20: Conference on Fairness, Accountability, and Transparency*.
[134] Maithra Raghu, Katy Blumer, Greg S Corrado, Jon M. Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. *ArXiv* abs/1903.12220 (2019).
[135] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct Uncertainty Prediction for Medical Second Opinions. In *ICML 2019*, Vol. 97.
[136] Naveen Raman and Michael Yee. 2021. Improving Learning-to-Defer Algorithms Through Fine-Tuning. *ArXiv* abs/2112.10768 (2021).
[137] H. G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. 2018. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics* 12 (2018).
[138] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* 35, 5-6 (2020), 413–451.
[139] J. Ren et al. 2023. Self-Evaluation Improves Selective Generation in Large Language Models. *Proceedings of Machine Learning Research (PMLR)* 239 (2023).
[140] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *IJCAI*.
[141] Ethan M. Rudd, Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. 2018. The Extreme Value Machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 3 (2018).
[142] Abulhair Saparov and He He. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *ICLR*.
[143] Advait Sarkar. 2023. Enough With "Human-AI Collaboration". In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*.
[144] Walter Scheirer, Anderson Rocha, Ross Micheals, and Terrance Boult. 2011. Meta-Recognition: The Theory and Practice of Recognition Score Analysis. *IEEE transactions on pattern analysis and machine intelligence* 33 (03 2011).
[145] Clayton Scott. 2012. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics* 6, none (Jan. 2012).
[146] Sambu Seo, Marko Wallat, Thore Graepel, and Klaus Obermayer. 2000. Gaussian Process Regression: Active Data Selection and Test Point Rejection. *Informatik aktuell* (1 2000).
[147] Burr Settles. 2009. Active learning literature survey. (2009).
[148] Glenn Shafer and Vladimir Vovk. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* (2008).
[149] Savyasachi V. Shah. 2024. Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records. *JAMA Network Open* 7, 8 (2024), e2425953.
[150] Irene Solaiman and Christy Dennison. 2021. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. In *NeurIPS*.
[151] Ricardo Sousa, Ajalmar Rocha Neto, Jaime Cardoso, and Guilherme Barreto. 2015. Robust Classification with Reject Option Using the Self-Organizing Map. *Neural Computing and Applications* 26 (01 2015).
[152] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the Black Box: User Experiences with an Inscrutable Algorithm. In *2017 AAAI Spring Symposium Series*.
[153] Tejas Srinivasan et al. 2024. Reducing Unnecessary Abstention in Vision-Language Systems via Evidence Seeking (ReCoVERR). In *Findings of ACL 2024*. https://aclanthology.org/2024.findings-acl.767.pdf
[154] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zheng Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chun-Yan Li, Eric P. Xing, Furong Huang, Haodong Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Sekhar Jana, Tian-Xiang Chen, Tianming Liu, Tianying Zhou, William Wang, Xiang Li, Xiang-Yu Zhang, Xiao Wang, Xingyao Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. 2024. TrustLLM: Trustworthiness in

Large Language Models. *ArXiv* abs/2401.05561 (2024).
[155] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*.
[156] Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. 2024. Learning to Defer to a Population: A Meta-Learning Approach. In *International Conference on Artificial Intelligence and Statistics*.
[157] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating Human + Machine Complementarity for Recidivism Predictions. *CoRR* abs/1808.09123 (2018). arXiv:1808.09123
[158] Stefano Teso. 2019. Toward faithful explanatory active learning with self-explainable neural nets. In *Workshop on Interactive Adaptive Learning*.
[159] Stefano Teso and Kristian Kersting. 2019. Explanatory Interactive Machine Learning. In *AAAI/ACM Conference on AI, Ethics, and Society*.
[160] Srikanth Thudumu, Philip Branch, Jiong Jin, and Jugdutt (Jack) Singh. 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data* 7, 1 (July 2020).
[161] Francesco Tortorella. 2004. Reducing the classification cost of support vector classifiers through an ROC-based reject rule. *Pattern Analysis and Applications* 7, 2 (2004).
[162] Rajeev Verma, Daniel Barrejón, and Eric T. Nalisnick. 2023. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. In *AISTATS 2023*, Vol. 206.
[163] Rajeev Verma and Eric Nalisnick. 2022. Calibrated Learning to Defer with One-vs-All Classifiers. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. PMLR.
[164] Edoardo Vignotto and Sebastian Engelke. 2020. Extreme value theory for anomaly detection – the GPD classifier. *Extremes* (2020).
[165] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014).
[166] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.* (2024).
[167] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017).
[168] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *CHI '20: CHI Conference on Human Factors in Computing Systems*.
[169] Zijie J. Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E. Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. 2022. Interpretability, Then What? Editing Machine Learning Models to Reflect Human Knowledge and Values. In *KDD*.
[170] Greta Warren, Mark T. Keane, Christophe Guéret, and Eoin Delaney. 2023. Explaining Groups of Instances Counterfactually for XAI: A Use Case, Algorithm and User Study for Group-Counterfactuals. *CoRR* abs/2303.09297 (2023).
[171] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.
[172] Zixi Wei, Yuzhou Cao, and Lei Feng. 2024. Exploiting Human-AI Dependence for Learning to Defer. In *ICML*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.), Vol. 235.
[173] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know Your Limits: A Survey of Abstention in Large Language Models. *Transactions of the Association for Computational Linguistics (TACL)* (2025), 529–556.
[174] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2021. Learning to complement humans. In *IJCAI*. Article 212, 8 pages.
[175] Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian J. McAuley, and Shuai Li. 2025. Beyond Chain-of-Thought: A Survey of Chain-of-X Paradigms for LLMs. In *COLING*. Association for Computational Linguistics, 10795–10809.
[176] Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Search-in-the-Chain: Towards the Accurate, Credible and Traceable Content Generation for Complex Knowledge-intensive Tasks. *CoRR* abs/2304.14732 (2023).
[177] Montreuil Yannis, Yeo Shu Heng, Carlier Axel, Ng Lai Xing, and Ooi Wei Tsang. 2024. Two-stage Learning-to-Defer for Multi-Task Learning. (2024).
[178] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS*.
[179] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A Survey of Knowledge-enhanced Text Generation. *ACM Comput. Surv.* 54, 11s (2022).
[180] Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, Tao Zhang, Chen Zhou, Kaizhe Shou, Miao Wang, Wufang Zhu, Guoshan Lu, Chao Ye, Yali Ye, Wentao Ye, Yiming Zhang, Xinglong Deng, Jie Xu, Haobo Wang, Gang Chen, and Junbo Zhao. 2023. TableGPT: Towards Unifying Tables, Nature Language and Commands into One GPT. *CoRR* abs/2307.08674 (2023).
[181] M. Zhang et al. 2024. Calibrating the Confidence of Large Language Models by Eliciting Fidelity. In *Proceedings of EMNLP 2024*. https://aclanthology.org/2024.emnlp-main.173/

[182] Xu-Yao Zhang, Guo-Sen Xie, Xiuli Li, Tao Mei, and Cheng-Lin Liu. 2023. A Survey on Learning to Reject. *Proc. IEEE* (2023).
[183] Michelle Zhao, Reid Simmons, and Henny Admoni. 2022. The Role of Adaptation in Collective Human–AI Teaming. *Topics in Cognitive Science* 17, 2 (Nov. 2022), 291–323. doi:10.1111/tops.12633
[184] Yuetong Zhao, Hongyu Cao, Xianyu Zhao, and Zhijian Ou. 2024. An Empirical Study of Retrieval Augmented Generation with Chain-of-Thought. In *ISCSLP*. IEEE, 436–440.
[185] Xiaojin Zhu. 2015. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29.
[186] Liu Ziyin, Zhikang T. Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2019. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems* (2019).
[187] Nikolas Zöller, Julian Berger, Irving Lin, Nathan Fu, Jayanth Komarneni, Gioele Barabucci, Kyle Laskowski, Victor Shia, Benjamin Harack, Eugene A. Chu, Vito Trianni, Ralf H. J. M. Kurvers, and Stefan M. Herzog. 2025. Human–AI collectives most accurately diagnose clinical vignettes. *Proceedings of the National Academy of Sciences* 122, 24 (2025), e2426153122. doi:10.1073/pnas.2426153122

Received 11 December 2023; revised 21 January 2026; accepted 26 January 2026