AI & SOCIETY
https://doi.org/10.1007/s00146-025-02401-y

OPEN FORUM

What it takes to control AI by design: human learning

Dov Te'eni[1] · **Inbal Yahav**[1] · **David Schwartz**[2]

Received: 9 February 2025 / Accepted: 19 May 2025

Abstract
Experts in government, academia, and practice are increasingly concerned about the need for human oversight in critical human–AI systems. At the same time, traditional control designs are proving inadequate to handle the complexities of new AI technologies. Incorporating insights from systems theory, we propose a robust framework that elucidates control at multiple levels and in multiple modes of operation, ensuring meaningful human control over the human–AI system. Our framework is built on continual human learning to match advances in machine learning. The human–AI system operates in two modes: stable and adaptive, which, in combination, enable the effective use of big data and the learning necessary for effective control and adaptation. Each system level and mode of operation requires a specific control-feedback loop, and all controls must be aligned for performance and values with the higher system level to provide human control over AI. Applying these ideas to a human–AI decision system for text classification in critical applications, we demonstrate how a method we call reciprocal human–machine learning can be designed to facilitate an adaptive mode and how oversight can be implemented in a stable mode. These designs yield high and consistent classification performance that is unbiased and closely aligned with human values. It ensures effective human learning, enabling humans to stay in the loop and stay in control. Our framework provides spadework for a model of control in critical AI decision systems operating in volatile environments, where humans continue to learn alongside the machine.

Keywords Human-AI system · Control · Reciprocal learning · Feedback · Oversight

1 Introduction

Leading experts in academia, government, and industry concur that establishing human oversight over critical AI applications is necessary to mitigate undesirable outcomes (Cohen et al. 2024; Grace et al. 2024). Effective human oversight would entail *monitoring* AI performance to detect any dangerous malfunctions or risky course of action and adjusting the AI performance accordingly. Nevertheless, with the growing complexity of AI models, it is not clear how monitoring and adjusting can be achieved without major breakthroughs to tackle challenges such as the need to design safety margins for dangerous AI capabilities, evaluate goal alignment with goals set by human designers, and incorporate interpretability and transparency necessary for effective adjustment (Bengio et al. 2024). Thus, humans' ability to control intelligent machines is questioned (Holzinger et al. 2025). Nevertheless, the need for control over AI becomes paramount when the complexity and uncertainty of unexpected machine behavior increase, as in cases where AI may alter designed behaviors or generate new behaviors when environmental changes make old behaviors inappropriate. In critical decision-making that threatens societal well-being, the risks and costs of erroneous decisions underscore the need for effective control over complex decisions. It is, therefore, crucial and urgent for society to determine what control over AI it wishes to achieve and how to achieve it. We do not attempt to predict whether AI will control humans or argue with those predicting when it will happen (e.g., Korbak et al 2025), but rather to determine what needs to be done if we desire meaningful human control. Accordingly, this paper presents a framework for analyzing the

✉ Dov Te'eni
teeni@tau.ac.il

Inbal Yahav
inbalyahav@tauex.tau.ac.il

David Schwartz
David.Schwartz@biu.ac.il

[1] Tel Aviv University, Tel Aviv, Israel

[2] Bar-Ilan University, Ramat Gan, Israel

Published online: 24 August 2025

desirability and feasibility of human control over AI used for critical decision-making.

One general solution to enable human oversight is to keep the human in the loop. Keeping the human-in-the-loop (HitL) has been advocated for several reasons: 1) improving performance due to the relative advantages of humans over machines, especially in dynamic environments, 2) maintaining human expertise through continual learning, building the user's trust in the machine, and 3) ensuring alignment with human values and being able to hold humans accountable. The reason for keeping HitL dictates the desired configuration of human AI and how it performs and adapts. For instance, the motivation for keeping HitL dictates the delegation of responsibilities between the human and the machine and, accordingly, what they should learn to be effective. This paper argues that ongoing learning is essential for controlling AI in complex and critical situations. Our motivation for keeping HitL is to enable human control and learning.

We build on the foundational work of Nobel Laureate Herbert Simon (e.g., Simon 1996) and a few of the thinkers he influenced, such as West Churchman, Russell Ackoff, and Karl Weick. Simon argued that control over artificial systems should not be regarded as a simple, one-way relationship but as part of a complex reciprocal relationship in which humans shape machines and machines shape society. Viewing a human working with an intelligent machine as a goal-oriented system, humans should ensure that the system aligns its performance with its client's human values. Moreover, the system and its subsystems must learn and adapt to operate effectively in a changing environment. Fearing that tight and instant control may stifle adaptation, Simon suggested that design should facilitate control through constraints and guidance. Thus, control in a system is necessary to warrant value alignment and effective adaptation, while also allowing the flexibility needed to adapt to changing contexts. As a result of the system's complexity and uncertainty, control must be designed as a multi-level mechanism distributed and coordinated throughout the system and its subsystems. Therefore, human control over AI can only be achieved if control is coordinated at all levels.

Simon recommends a combination of regulation and design to facilitate control. In our paper, we focus on design and revisit its connection with regulation in our discussion. We present a design methodology to ensure adequate human oversight for critical AI applications. The design methodology for human–AI systems is grounded in Simon's concepts of learning and adaptation, value alignment, and control in the face of uncertainty and complexity, as well as subsequent extensions, including trusted feedback loops and meaningful control, which enable accountability. These ideas are pertinent to critical decision-making in a dynamic environment. We begin with an example that demonstrates these ideas.

Consider an AI-based decision-making application for detecting suspicious messages on social media, overseen by a human cybersecurity expert (the design of this application is described in Sect. 2.2). The AI comprises several machine learning classification models trained on previous judgments of suspicious messages. The expert is responsible for detecting changes in the environment and adapting the AI models accordingly. The cost of error in detecting suspicious messages is considerable. The application designer worked with the assumption that human oversight is essential, but that it is feasible only if human experts continually develop and maintain the requisite capabilities to detect and correct erroneous decisions made by AI. Without a deep and up-to-date understanding of the phenomenon and its environment, the expert will be unable to evaluate and correct the AI classification models and, importantly, may feel unconfident about overruling AI.

In the scenario we develop below, once AI is trained, it operates autonomously, capitalizing on its ability to make decisions immediately based on big data. At the same time, the human is kept on the loop for oversight purposes only. Nevertheless, when the underlying decision environment changes or the decision probability distribution changes beyond a certain threshold, the mode of operation shifts back to a learning session in which the machine and the human expert learn from each other reciprocally. In this mode of operation, humans are kept in the loop to reinforce machine learning and learn from it themselves. In addition, any adaptation to the system based on learning should be controlled to test whether value alignment persists. For example, the expert may want to confirm that AI-based classification does not unfairly discriminate against specific populations. The expert must also learn how to oversee the machine to ensure it is aligned with human values. The learning session continues until it reaches a point of saturation. The AI can then resume autonomous operations until the next learning session. Importantly, different modes of operation will require different control designs, as we argue below.

As mentioned above, the high-level relationship of 'human control over AI' must translate into a multi-level structure of controls. In systemic terminology, overall control of system performance entails control mechanisms at both the system level and the lower levels of the subsystem. An example of a delegation of responsibilities is that the human expert controls when to enter a learning session, and the machine controls the operation of its classification models, but both controls are coordinated. Thus, we examine the design requirements of control at different levels and modes to ensure they align with the overall control requirements.

When examining the relationship between humans and AI as the basis for designing control, we note that Simon's perspective is not the only one on the human–AI relationship and that different perspectives dictate different

designs. Designers delegate responsibilities between humans and machines, formulate goals and values, and justify the design's working assumptions according to their design perspective (Suchman 2007). The delegation of responsibilities necessarily affects the locus of control. For example, to leave control to humans, designers may choose not to delegate opaque decisions to machines (Robbins 2025). Others may be tempted to harness AI capabilities to achieve better-than-human decision performance subject to pre-designed controls (Bommasani et al. 2021), despite widespread warnings that superhuman intelligence may lead to disastrous outcomes for society (Sparrow 2024). Our approach aims to achieve better decision-making performance with human oversight while enhancing human capabilities through continual learning.

We develop our arguments in the context of human–AI systems designed to enhance decision-making. We regard human–AI decision systems as Singerian inquiry systems. Singerian inquiry systems extend Simon's view of a system to emphasize continual learning through dialog between participant decision-makers, who entertain multiple perspectives, and through value-laden inquiries that question the boundaries and assumptions of the system (Churchman 1971). *Human–AI decision systems* combine human judgment and artificial intelligence capabilities to produce effective decision outcomes by leveraging the strengths of both humans and machines, ensuring oversight, adaptability, and accountability in complex decision environments. We, therefore, conceptualized learning in the context of complex decision-making as a process of making sense of a dynamic world (Weick et al. 2005).

Our level of analysis is a human–AI decision system. However, we believe it can serve as a basis for examining the control of AI more broadly when examining a collection of many systems. We restrict the analysis to critical decision systems in which the needs and preferences of humans supersede those of the machine; therefore, the system must serve humans and ultimately leave them in control.

The following section describes systems theory as it applies to a human–AI system, and the third section presents a framework for controlling AI. Applying the framework, the fourth section offers a design of a human–AI system with autonomous machine classification. It demonstrates how designs can control AI with appropriate feedback that supports control and learning. The feedback is generated using a recently developed computer-based method, which relies on reciprocal human–machine learning. We use this method to demonstrate the feasibility of applying our framework. The final section generalizes from our case of human–AI systems for message classification to decision-making more generally. It acknowledges the need for both regulation and design to warrant human control.

2 Systems theory

Compared to traditional human–computer interactions for decision-making, human interaction with intelligent machines requires new ways of conceptualizing these interactions to cope with the higher complexity and variety of decision tasks that AI can now perform (Ågerfalk et al. 2022). Indeed, researchers have proposed various metaphors to conceptualize human–AI configurations, including a tool (Shneiderman 2020), an agent (Harari 2024), a human–machine team (Tsamados et al. 2025), and a collaborative partnership (Woods & Hollnagel 2006). Underlying all these configurations of humans and machines is the assumption that machines can perform some decision-making elements better than humans or vice versa. Nevertheless, the criteria for delegating responsibilities between humans and machines and how the delegation is determined and controlled differ from one configuration to another. The growing capabilities of AI to perform more elements of decision-making effectively complicate the delegation of responsibilities compared to traditional models of human–computer interaction, and, in particular, introduce new criteria, such as moral considerations, which were generally absent in past models of task allocation in human–computer interaction (Te'eni 2025). Following other scholars, such as Russell (2022) and Baird and Maruping (2021), we use the term "responsibility delegation" rather than "task allocation" when examining work in human–AI systems, emphasizing the complexity and value-laden nature of goal-oriented assignments to intelligent agents.

We chose Simon's systemic approach, with its rich theoretical insights into the role of control and its wide range of applications that involve decision-making. Simon begins his analysis of the artificial (designed by humans) with the realization that humans are limited by their ability to process information, an assumption he labeled 'bounded rationality'. AI, by comparison, can process enormous quantities of data quickly and make informed decisions. Rejecting the tool metaphor of human–computer interaction, Simon regards the artificial and the human working together towards specified goals as a system in which one shapes the other. The artificial and the human can be seen as subsystems, each performing the responsibilities delegated to it in coordination with the other subsystem. Control is necessary to ensure the effective operation of a system aligned with the human client's values.

Below, we expand on Simon's systemic approach. First, we apply systems theory to help determine which aspects of human–AI systems that learn and adapt can be controlled. Second, in Sect. 2.1, we utilize systems design principles to determine how to achieve control through feedback.

2.1 Systems structure and roles

In *The Sciences of the Artificial*, Simon builds on *systems theory* to analyze the form and function of artificial systems in relation to their environment (Simon 1996). We follow suit using the terminology and principles of systems theory developed by Churchman (1968) to examine the human and the AI as a system composed of human and artificial subsystems. A system has a goal and a measure of performance that reflects the effectiveness of achieving the goal and the efficiency of utilizing its resources to achieve the goal. It operates within an environment that constrains its operations and affects its performance measures. The system comprises goal-oriented subsystems capable of performing specific responsibilities that, in coordination, co-produce the system's performance measures. A system has a client, a system manager (Churchman uses the term 'decider), and a designer. The system manager delegates responsibilities and resources to impact performance. The designer determines the structure of the system, its subsystems and their relationships, the responsibilities and the criteria for delegating them among subsystems, and the rules of operation and incentives. The goal-oriented system is designed to serve the system's client, aligning the system's performance measures with the client's values. The designer is assumed to strive to maximize value to the client sincerely; additionally, a built-in mechanism warrants that the intended design is realized and its outputs are valid, e.g., through a set of validation rules (Churchman 1971).

The system serves its clients by achieving appropriate endogenous and exogenous goals with respective performance measures. Throughout this discussion, we assume that the client is human and that the system should align with the client's stated values. Values are the principles and priorities of the system's clients that underlie its purpose (Churchman often uses the term 'wants'). Moral considerations are a significant part of human values, which evolve through continual learning and ethical reflection. Adhering to the client's values can be viewed as an endogenous goal, sometimes called purpose, in contrast to exogenous goals. Values may compromise exogenous goals. In our example of detecting suspicious messages, an exogenous goal is correctly detecting all cybersecurity threats, and a possible measure of performance is the average accuracy of the classifications. An endogenous goal could be maintaining a fair and unbiased classification algorithm at the expense of lower accuracy.

People today play multiple roles in their interactions with AI (Baskerville and Myers 2023). First, as a client, the human–AI system serves the human. Second, coordination among all components of the human–AI system requires a human system manager who can realign goals and change decisions accordingly. These management capabilities include delegating and re-delegating responsibilities for specified goals to the machine and controlling their implementation to ensure the machine does not pursue goals misaligned with those of the human client.

In addition to the system roles of client and system manager, humans can also play the role of the system's designer, who conceptualizes the system in such a manner that it leads to outcomes that optimize the system's value to its clients (Churchman 1968). It is assumed that the designer's intentions are aligned with the client's interests, and furthermore, that the system is controlled to confirm that the designer's intention is realized. The paper does not address aligning the goals of AI designers and clients, e.g., laws enforced by governments that ensure products do not violate consumers' rights.

2.2 Systems operation

Adaptive systems can operate in alternative modes and adapt their goals and delegation of responsibilities accordingly. A system can operate in a *stable mode*, aiming to perform effectively while maintaining its current configuration, or in an *adaptive mode*, which seeks to adjust its configuration. von Bertalanffy (1968) highlighted the crucial role of the adaptive mode in enabling the stable mode's operation and control. In contrast to stable mode, adaptation requires learning and making sense of the environment, particularly detecting and interpreting changes within the system and in its environment to realign goals and revisit system performance. The adaptive mode becomes crucial when system complexity and environmental uncertainty are high. The delegation of responsibilities requires corresponding decisions on allocating resources between human and AI subsystems and depends on the system's mode of operation. In our example, humans implement learning-related goals in adaptive but not stable modes.

"Learning is any change in a system that produces a more or less permanent change in its capacity for adapting to its environment." (Simon 1996, p. 100). Churchman (1968) elaborates that an adaptive system should rely on continual learning by its subsystems using their local intelligence to sweep the environment. Still, he warns that learning in subsystems must be controlled to avoid misalignment between the goals of the adapting subsystems. Moreover, controlled adaptation should include continual value-laden reflection to avoid morally adverse outcomes. These principles of learning and adaptation play a crucial role in our design approach, where the human and AI subsystems, each with their respective learning capabilities, make sense of the environment and learn from one another.

Concentrating on human–AI systems for critical decision-making, we adopt Weick's sensemaking model of learning to make decisions. Weick et al. (2005) proposed

an Enactment–Selection–Retention model to describe how agents make sense of their environment and decide accordingly. People enact, rather than only react to, their environment, select an appropriate concept or interpretation to make sense of phenomena, and retain successful concepts for future action. Colville et al. (2016) extend the concept of sensemaking in environments characterized by dynamic complexity to emphasize the importance of learning from feedback that challenges existing knowledge and guides the adaptation of future actions. According to this model, learning by an agent in a human–AI system involves iteratively perceiving and experiencing events in the environment or within the system, making sense and drawing inferences in light of extant knowledge, relying on feedback and experimentation, and developing new knowledge structures that may impact current and future decisions. As agents can be human or artificial, learning outcomes are represented as knowledge that is accessible to both. Such learning requires time and effort, proper conditions for learning and experimentation, feedback and control, and effective retention. These conditions guide the design of the adaptive mode of operation dedicated to learning and adaptation. Moreover, due to the differential learning capabilities and limitations of humans and machines, they will require different conditions, which are elaborated upon in Sect. 2.2.

3 Control

The systems theory detailed above is the basis for analyzing control. We present the analysis in three steps. First, we conceptualize control as the monitoring and adaptation of system performance, building on Simon's notion of coordinated multi-level control. Second, we extend performance considerations to moral considerations in control. Third, we address the complementary role of trust in the feedback that supports control.

3.1 Complex systems require a variety of controls

Bounded rationality in systems that face a variety of uncertain events in their environment leads to complexity (Simon 1996). Complexity is often controlled by decomposing hierarchical systems into subsystems and then subdividing these subsystems into sub-subsystems, enabling a multi-level control mechanism. Each goal delegated to a human or machine is controlled with a corresponding feedback loop designed to fit the system's goal and source of uncertainty (e.g., internal or external), the subsystem (human or machine) achieving the goal, and the mode of operation (stable or adaptive). Human control over AI at the system level will be effective only if lower level controls ensure that all subsystems coproduce the system's performance measures, including the coordination between the human and machine subsystems in both modes of operation.

Controlling system performance in the face of uncertainty ensures that the system continuously pursues its specified goals and does so efficiently. Ashby's famous law of requisite variety states that "The variety in the control must be equal to or greater than the variety in the disturbance." Ashby (1956, p. 206) is popularly referred to as "Only variety can destroy variety." A more significant number of sources of uncertainty necessitates a more flexible system that can effectively control disturbances through appropriate means. Nevertheless, too much variety may increase complexity to a level where control becomes inefficient.

The source of uncertainty in human–AI systems can be either internal to the system or external to its environment. Internal uncertainty can arise from the human, the machine, or the interaction between them. In traditional human–computer interaction (non-AI), human error due to the user's limited cognitive capacities, attitudes, and biases was expected, and controlling for error was a common design goal, often involving corrective feedback (Reason 1990). The introduction of AI has intensified the need to control errors and deviations from expected behavior that originate within the AI models used and in their interaction with the environment. Unexpected behaviors of AI models due to randomness in the models have been noted, especially with generative AI, and various guidelines have been proposed to control deviations (OWASP 2023). In addition, the increased complexity that AI presents to the user amplifies the impact of uncertainty and the corresponding need for control (Simon 1996). It has become more challenging to monitor AI performance, and, moreover, it has become especially difficult for the user with bounded rationality to adjust the AI. Thus, the adaptive mode will require support in learning to be effective.

Similarly, external uncertainty due to unknown changes in the environment necessitates control to identify when to adjust the system's operation and goals to prevent detrimental outcomes or minimize the probability and cost of error. While the design of monitoring depends mainly on the volatility of the environment, the need for adjusting the AI depends on both external uncertainty and the complexity of the AI, as well as internal uncertainty. Therefore, control in the face of external uncertainty also requires learning and support to overcome the limitations of bounded rationality.

Control has similar but distinct purposes in the system's two modes of operation, implying different forms of human control. These different forms of control in the stable and adaptive modes of operation can be referred to, respectively, as human-on-the-loop and human-in-the-loop (see descriptions of these terms in Robbins 2025). In stable mode, control and its corresponding feedback are designed to self-regulate the system through negative feedback or alert

the system manager to change modes. Human control is ***on*** the loop, overseeing operations. In adaptive mode, control directs the system to transition to a new state (morphogenesis) based on developmental feedback. Human control ***in*** the loop controls operations and supports learning.

Figure 1 illustrates control at two levels: one at the system level and the other at the subsystem level, operating in both stable and adaptive modes. Human control at the system level oversees the entire system's operations and the transitions between different modes of operation. While internal and external sources of uncertainty require control in both modes, uncertainty stemming from environmental changes is typically the primary trigger for transitioning from stable to adaptive mode, especially in non-generative AI, where the (often limited) computational context space may not be regularly updated to reflect changes in the environment. The various roles of control must be aligned to achieve overall control by humans over machines (performance alignment). The chain of control is no stronger than its weakest misaligned link. For instance, humans may control performance in adaptive mode but not the transition from stable to adaptive, in which case, human control over AI is diminished.

Churchman (1971) emphasized the importance of value-laden reflection to confirm that moral considerations are incorporated into decision-making, as well as the efficiency and effectiveness considerations of control, as explained earlier. Similarly, the role of learning in controlling performance is crucial, as without acquiring new or updated moral concerns, the system may not accurately reflect the client's current values (Ackoff 1974). In Fig. 1, the moral considerations are defined as additional internal purposes. In the message detection example, an internal purpose could be to prevent unfair classifications resulting from a biased data distribution.

Moral considerations are paramount when discussing human control over AI (Siebert et al. 2023). From a moral perspective, the concept of control must explicitly include the ability to warrant machine-autonomous behavior that complies with human moral principles. Initially coined in the context of autonomous weapon systems, the term "meaningful human control" (Robbins 2024; Santoni de Sio and Van den Hoven 2018) connotes that humans must control and be *morally* responsible for the performance of autonomous agents. In addition to the ability to monitor and react to deviations from moral standards and relevant environmental changes, meaningful human control must also be able to trace back the outcome of the system to at least one responsible human decision-maker.

3.2 Degree of control

Control is not binary. Simon warned against control that may be too tight, as it restricts the flexibility needed for adaptation, arguing instead for control through incentives and constraints that guide behavior. By the same token, controlling AI begs the question of to what degree. For instance, guiding generative AI through more detailed prompts or directly reducing its level of algorithmic randomness may reduce the risks of errors but affect the innovation in its responses.

Trustworthy AI and controllable AI have been regarded as complementary (Schoeller et al. 2021) or alternative approaches to address AI dangers (Kieseberg et al. 2023). Undoubtedly, the relationship between control and trust is complex. A high-level expert group on AI recently defined

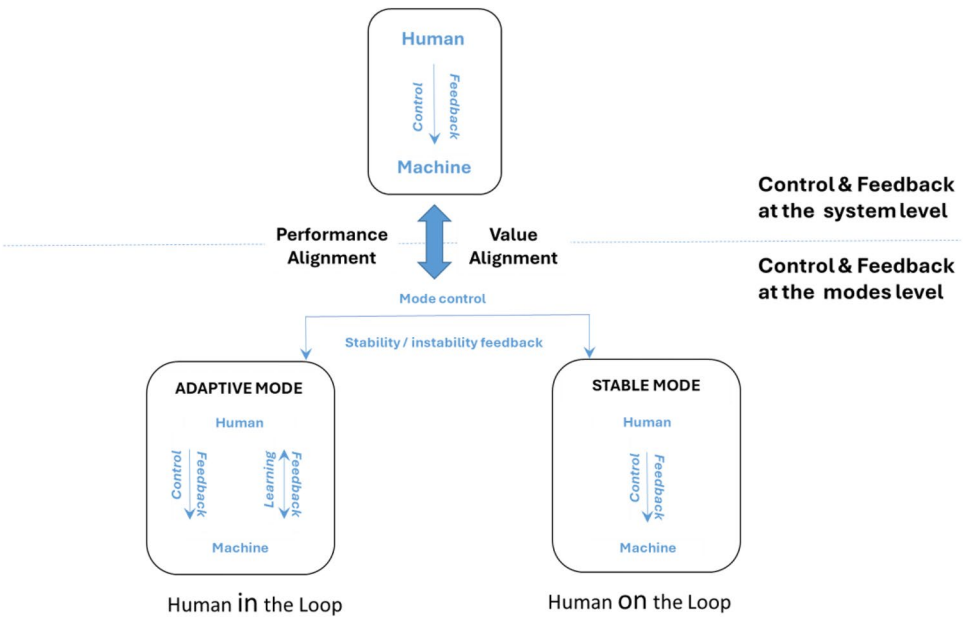

Fig. 1 Control and feedback in the face of uncertainty at the system level and the modes level

the requirements of trustworthy AI, the first of which is human agency and oversight:

> "AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms must be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches" (European Union 2025).

A system manager who trusts the system (or subsystem) may be satisfied with a lower degree of control. For that to happen, the system manager must trust the feedback loops that indicate the system's state, quality of decision processes and outcomes, and alignment with goals and values. Therefore, the design of control and feedback in human–AI systems will need to consider ways of increasing trust.

To summarize the discussion of variety and degree of control, Fig. 1 illustrates a variety of controls, each designed specifically to fulfill its particular role. In contrast, all controls are coordinated to align with the overall system's goals. To paraphrase Ashby (1956) in his description of the law of requisite variety, "Only variety can control variety." Furthermore, each control mechanism in each role will require a distinct form of feedback to ensure effective goal-oriented performance, as elaborated in the following subsection. Finally, control and feedback in human–AI must be trustworthy.

3.3 Feedback design

Feedback enables control, allowing systems to function in stable and adaptive modes and achieve their internal and external goals (Churchman 1968). In particular, control directs the response needed, and feedback indicates how to adjust accordingly. The design of feedback in human–AI systems is challenging for several reasons. First, to enable control, feedback must be tailored to the specific purpose of the control and mode of operation. For instance, feedback that allows control to achieve internal goals may differ from feedback that enables control to achieve external goals. Each feedback design must be tailored to effectively serve the recipient, monitor, or adapt to the specific goal. Second, in addition to its central role in enabling control, feedback can also support learning, provided it is designed accordingly. In Fig. 1, we added a learning-feedback loop in adaptive mode to highlight the distinct form of this feedback. Third, designing bi-directional feedback between the human and machine subsystems in adaptive mode is especially challenging because it involves feedback tailored for two distinct learning processes: human and machine learning. Feedback designed to serve a human differs from that designed to serve a machine due to the distinct information processing styles and capabilities, as well as the distinct functions the feedback plays.

In the stable mode of the human–AI system, the AI makes decisions and solves problems autonomously, with integral oversight as part of the system's design. Oversight requires automatic monitoring and control feedback to the human that signals when decision-making is no longer effective. At that point, autonomous decision-making must be aborted, and the mode of operation shifted from stable to adaptive mode. Monitoring autonomous decision-making is complicated when there is no established desired norm (a 'ground truth') against which to compare actual performance as a basis for generating the appropriate feedback. One solution is to monitor the AI's confidence level (Gal and Ghahramani 2016; Lakshminarayanan et al. 2017) and automatically transition from stable to adaptive mode when the confidence falls below a certain threshold. Threshold in this context can refer to a change in concept drift—the underlying data distribution (Greco et al. 2025), a change in estimated prediction confidence (Kivimäki et al. 2025), or even after a specific period of time (Bodor et al. 2023).

Feedback in adaptive mode is more complex. When the system's operation complexity is high, as assumed, autonomous decision-making cannot be easily adjusted based on negative feedback alone without evaluation and learning (Simon 1996). Such learning typically requires supervision by a human expert. Without understanding why and how to adapt the decision-making mechanisms of the AI, it is impossible to ensure that the adapted system will stay on course towards the goals set by the designer. Therefore, the adaptive mode requires two different forms of feedback: feedback that supports the *control* of learning to decide when to stop learning and transfer to stable mode and feedback that supports the *learning process* by evaluating the decisions made (our discussion does not address a third possible role for feedback, namely, motivation (Te'eni 1992).)

Systems thinking raises another challenge for feedback design: feedback to humans must be trustworthy. Trust between agents typically stems from mutual understanding. A typical design principle is to strive for the understandability of AI systems by making their algorithms more transparent (Rudin 2019) or providing explanations of outputs (Ribeiro et al. 2016), even though we lack conclusive evidence on the value of explainability (Wood 2024).

Figure 1 depicts our framework of control and feedback, as discussed in this section, through the lens of systems theory. A system will typically require a variety of controls with several possible roles, including (a) control to ensure performance is effective and efficient, (b) control to ensure learning is effective, (c) performance alignment between subsystem and higher level goals, (d) value alignment between subsystem behavior and higher level moral considerations, and (e) control over shifts from one mode

of operation to the other. Every system (or subsystem) goal requires a dedicated control-feedback loop, and each control-feedback loop necessitates a tailored design to fulfill its role effectively.

In the next section, we demonstrate the feasibility of this model (Fig. 1) with a concrete application around the cybersecurity detection example we introduced earlier. In transitioning from analysis to design, pragmatic considerations emerge that lead to 'satisficing' decisions (Simon 1996), i.e., satisfactory rather than optimal solutions for a realistic world, which is far more complex than the theoretical model presented above. It may not be feasible to control entirely and all the time, and it may be necessary to operate with acceptable levels of risk. We return to this issue in the discussion. For now, we note several such considerations that are not visible in Fig. 1 but are, nevertheless, important in moving to the application. Higher complexity and uncertainty require greater variety, a higher degree of control, and more complex designs of corresponding feedback. Nevertheless, a high degree of control may stifle innovation by limiting the flexibility needed for adaptation. Trusted and understandable feedback may compensate for a lower degree of control but may result in less efficient outcomes. The challenges in designing control, particularly in tailoring effective feedback, are significant, as discussed in the next section.

4 Design of a human–AI classification system

We designed and implemented a human–AI classification system, called Fusion (Zagalsky et al. 2021), based on the functional requirements determined by applying our design framework (depicted in Fig. 1). This section aims to demonstrate the framework's feasibility and ability to identify the fine-tuning required in designing different feedback loops. The human–AI decision system operates in stable and adaptive modes following our design guidelines. It combines control over autonomous message classification in stable mode and reciprocal human–machine learning (RHML) in adaptive mode. Alternating between adaptive and stable modes allows, respectively, continual learning and efficient classification of big data. Continual learning is necessary when the content and form of messages change continuously and when classification knowledge must be updated accordingly. Figure 2 illustrates the two modes and their interaction.

Returning to the example in the Introduction, consider a configuration of a cybersecurity expert and an AI-based classification model that must classify (decide) whether an incoming message is a cybersecurity threat. In stable mode, messages are classified automatically by a machine learning (ML) classification model that has been previously trained, with human experts overseeing the classification process. In adaptive mode, the human and the machine classify messages, and each learns from the other through the feedback they receive. Using screenshots from *Fusion*, we demonstrate the design of the cybersecurity case with various feedback screens. We begin with the adaptive mode.

4.1 Adaptive mode: reciprocal human–machine learning (RHML)

For the adaptive mode, we propose RHML as a method for building a human–AI decision system and its subsystems capable of learning with their 'local intelligence' directly from the environment and, in parallel, capable of coordinating adaptations and re-aligning goals. Moreover, RHML allows each subsystem to leverage its learning capabilities while allowing for learning feedback from one subsystem to another, thereby boosting mutual learning and capitalizing on the relative advantages of each subsystem.

In our example, human experts make sense of new cases (messages on an online forum) in light of their accumulated knowledge and pass their insights as feedback to the machine. Meanwhile, the machine provides feedback to the experts on their classification performance and consistency. Thus, in our example, the human–AI system combines AI-based text classification algorithms with human sensemaking processes to detect imminent cybersecurity threats in text messages. The adaptive mode in Fig. 2 depicts two feedback loops, one for human learning and

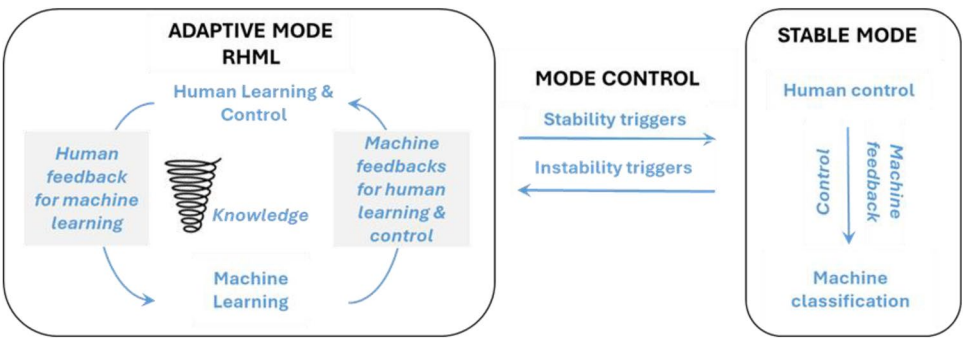

Fig. 2 Reciprocal human–machine learning for adaptive mode, and human control over AI for stable mode

the other for machine learning. Both loops are linked to knowledge representations used by humans and machines alike. The second feedback loop aligns with current trends in reinforcement learning, utilizing human feedback (Lin et al. 2020). In addition to classifying and learning from feedback generated by the machine, the human must control reciprocal learning and decide when to stop learning.

While machines and humans learn differently, they can learn from each other, and furthermore, they can prompt each other to reflect and reassess (Simon 1996). In RHML, we follow Churchman's idea of taking others' points of view into reassessment, as eloquently stated by Churchman: "The systems approach begins when first you see the world through the eyes of another" (Churchman 1968, p. 231). Indeed, learning by juxtaposing alternative perspectives is a particularly effective form of learning. In our case, the feedback from the machine to the human presents a different point of view on the same information. Still, the complexity of communicating perspectives requires appropriate designs to confirm the mutual understanding necessary for solving problems, analyzing, explaining, and judging through repeated and intertwined learning processes. Our proposed feedback designs, shown below, aim to address these challenges by incorporating two feedback loops, facilitating an iterative learning process.

In our studies of using *Fusion*, human experts established a small, labeled dataset (~ 500 messages, of which approximately 2.5% were labeled as imminent cybersecurity attacks) to train the machine learning (ML) classification models. Once trained, the ML classification model system engages with a domain expert in several cycles of reciprocal learning. The machine classifies multiple messages in each cycle and provides feedback, analyzing both aligned and misclassified cases. The expert then learns new considerations for future message classification from this feedback and shares these insights with the ML system for further refinement. The machine updates its classification models and enters a new cycle to classify new content. Based on the feedback, the expert then develops new criteria for classifying messages in the future and shares these insights with ML for further enhancement. The machine modifies its classification models and starts a new round to classify new messages. These rounds of mutual learning continue until a point of saturation is reached, at which the human decides that new feedback from the machine adds little.

We experimented with two classification models: Reflection and New Insights. *Reflection* was an unsupervised model reflecting human knowledge, using a sentiment scoring-based algorithm to score messages' risk (Pang and Lee 2004). *New Insights* had two main components: enhancing the knowledge based on semantic similarity and classification using word2vec (Mikolov et al. 2013) to represent the textual input and lasso regression via glmnet (Friedman et al. 2010), with higher attention given to human knowledge.

The feedback loop from the machine to the human must provide the relevant information to enable effective learning and provide it in an understandable form. Figure 3 is a screenshot from Fusion. The left-hand part of the screen shows the menu options, and the lower part shows machine-generated feedback to the expert user. Fusion supplies two types of learning feedback: outcome (performance) and explanatory. The outcome feedback is presented as performance metrics, including area under the curve (AUC), recall, precision, accuracy, and a confusion matrix, which are generated at each iteration.

The explanatory feedback focuses on the reasoning behind classifications, which refers to the process of arriving at the classification outcome. It often "drills down" to specific cases to examine the rationale for classification or reveals differences between human and machine rules and lexicons. Figure 4 is one of several screens showing explanatory feedback. Four classified messages are shown in the context in which they were communicated. Messages are organized according to the confusion matrix, and the four shown are from the highlighted group of ten messages correctly identified as suspects (true positives). The (red and blue) underlined words are terms that increase the model's propensity to classify a message as suspect or non-suspect. The marked word "Exploit" is shown in its linguistic context, that is, the words surrounding it in the message. The marked word can also be seen in its context across messages through another function. By hovering over the term, the user can see its propensity to classify the message and the other lexicon terms that affected this propensity. In addition, the interactive category tree (depicted on the right side of Fig. 4) indicates the term's location in the expert's lexicon.

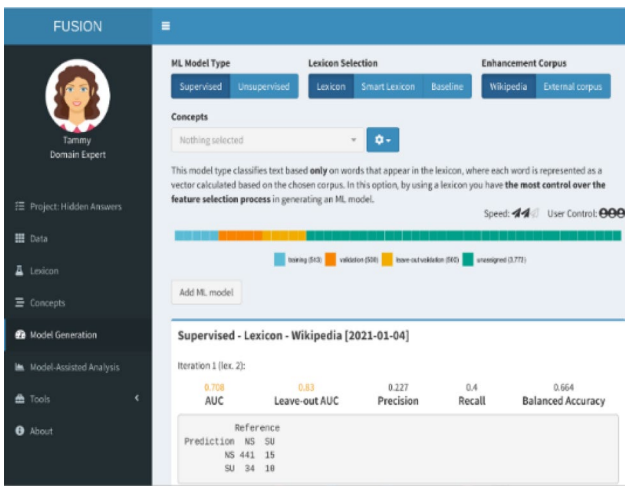

Fig. 3 Fusion screen showing menu, model parameters, and outcome feedback (adapted from (Zagalsky et al. 2021))

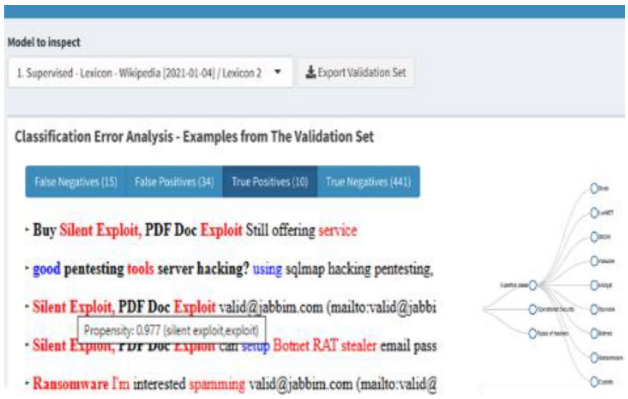

Fig. 4 Drill-down feedback on specific case (taken from (Zagalsky et al. 2021))

The tree allows the user to view the term in its situational context.

To make the machine-generated feedback understandable, we used available software that simulated explanations of the machine's rationale for its classifications. Initially, we implemented the Local Interpretable Model-Agnostic Explanations (LIME) software package to support explainability (Ribeiro et al. 2016). While LIME remains widely used, it represents just one approach within the rapidly evolving field of explainable AI (XAI) (for a recent survey, consider Bennetot et al. 2024). In our implementation, as our corpus grew, the LIME technique became impractical; therefore, we adapted it to approximate each term's propensity by averaging its prediction impact across instances. Feedback can then be tailored accordingly, as demonstrated in Fig. 4. The explanatory feedback from *Reflection* reflected on the knowledge provided by the expert, specifically the propensity of each corpus term to predict suspect messages. *New Insights* allowed us to suggest new knowledge to incorporate.

The possible tradeoff between explainability (or interpretability) and accuracy is a debated dilemma. On the one hand, the transparency of AI models has been recognized as a design goal to enable human oversight, promote trust, and facilitate accountability (Bengio et al. 2024). On the other hand, requiring explainability may limit the adoption of more accurate models (London 2019; Robbins 2025). Designers will need to explicitly resolve the tradeoff in light of the design goals of their particular project, rather than adopting a blanket policy. In our project, explainability was necessary to enable effective human learning.

Joint knowledge representations are the basis for mutual understanding. Part of the joint knowledge representations is the expert's accumulated classification knowledge, stored as traceable cognitive maps and an extended lexicon—a lexicon comprising terms, phrases, patterns, and rules (Te'eni et al. 2023). Lexicon generation employed qualitative data analysis techniques Silverman et al. (2025) to generate knowledge. Another form of representation is cognitive maps. The cognitive maps serve as the basis for constructing the second feedback loop, which extends from the expert to the machine.

Our studies found that experts were more likely to detect unique, non-recurring environmental changes that required learning and adaptation compared to the AI-based models we used. Enhancing decision-making by transferring experts' knowledge of these changes to machine models aligns with the systems theory principle of coordinating adaptation between subsystems and delegating responsibilities based on the relative advantages of each subsystem.

Aggregate indicators of the learning were computed to support control in RHML. Based on the aforementioned measures of AUC and complexity derived from the cognitive maps, learning trends were computed and used to inform the human decision on when to abort learning and re-enter stable mode. In our experiments, after eight iterations, the machine and the human reached an agreement on mutual knowledge, theoretically allowing the system to transition into a stable mode.

4.2 Stable mode: autonomous decision-making with human oversight

In stable mode, the augmented classification models operate autonomously. Our initial implementations partially automated control through the use of probabilistic thresholds. These thresholds are based on the distribution of predicted probabilities, where a significant shift signals a potential environmental change. Such shifts may indicate an increase in attack patterns, changes in language cues, or inaccuracies in previously learned knowledge. When these thresholds are triggered, the human enters the loop to detect situations where autonomous classification should be aborted and reciprocal learning re-established. The human expert is expected to monitor the environment for meaningful changes or apparent misclassifications reported from outside the system, e.g., by other agents encountering security breaches.

As explained, the system can provide the user with probabilistic indications of classification accuracy in the form of AI confidence scores. Suppose the confidence scores fall below the predetermined threshold or the overall distribution shifts significantly. In that case, the system alerts the human expert, who decides whether to regain control and transfer to adaptive mode. In dynamic and turbulent environments, the human–AI decision system remains in stable mode for relatively short periods and moves to adaptive mode more frequently, requiring tighter controls. In any event, humans acting as system managers will have a 'red' button to transfer control at their discretion immediately.

5 Discussion

Ensuring meaningful human control over AI in complex decision-making systems requires appropriate designs to keep humans in and on the loop. We propose a design framework rooted in the principles of systems theory in which learning plays a crucial role. According to these principles, overall system-level control entails aligned control at lower levels, and complex decision-making requires continual learning for effective control and adaptation. The design framework emphasizes the importance of integrating robust control mechanisms with adaptive learning processes to manage internal and external uncertainties effectively. The human–AI decision system operates in stable and adaptive modes, enabling both effective control and efficient AI performance on big data. Pragmatically, the framework identifies the types of controls required for a human–AI decision system, and the specific application demonstrates its feasibility while raising the challenges facing designers. We now address these challenges in light of our framework and findings.

5.1 Systems theory and learning

Applying systems theory to examine a human–AI system for critical decision-making resulted in several broad guidelines. First, control should be operationalized as a multi-level set of coordinated control mechanisms, each tailored to a specific goal. The hierarchical decomposition of a system into subsystems and modes of operations guides the design of human control over AI. Second, each control mechanism is designed to support monitoring and adjusting, and dictates corresponding feedback. Third, the two modes of operation, with their dedicated control-feedback mechanisms, require additional control to coordinate the transitions between these modes. Fourth, human control should be operationalized to monitor and, accordingly, adjust the system to improve and align with goal-oriented performance measures, human values, and learning effectiveness.

The two modes of operation are designed to support decision-making in dynamic environments where knowledge may need periodic adjustments, and the shifting decision from stable to adaptive mode can tolerate the delays that occur from threshold-based decisions. It may be the case that environmental changes do not entail acquiring new knowledge, as existing sensemaking processes remain applicable. In those cases, a single stable human-in-the-loop mode may suffice. On the other extreme, when the environment requires frequent adjustments to knowledge or values, or when the risk of false positives is too high, a single mode in which decisions are always made jointly may be more suitable.

Delegation of responsibilities is central to systems theory and is traditionally determined by the system manager seeking to optimize performance subject to system constraints. Delegating responsibilities between humans and intelligent agents is considerably more complex when considering human control, the understandability of AI algorithms, trust in technology, and accountability. In the design of controls for our message classification case, we assumed that control in stable mode is delegated to the human expert, while detecting suspected messages is delegated to the machine. Based on this delegation, we design specific controls coordinated to co-produce overall control over the system. We do not delve into the assumptions and mechanisms for determining what to delegate to whom, as this is a complex topic that deserves a dedicated analysis. Work on delegation to intelligent agents, such as Russell (2022) and Baird and Maruping (2021), will become an important basis for designing controllable and trustworthy AI.

5.2 Feasibility of control and learning

While there is growing agreement about the desirability of controlling AI in critical areas, there is substantial concern about the feasibility of doing so. Holzinger et al. (2025), for example, question the feasibility of meaningful human control in critical domains such as new biotechnology. They suggest that techniques such as keeping HitL, explainable AI, and regulation may not be capable of providing complete control over increasingly sophisticated 'black box' AI, particularly in complex decision environments. They believe that new frameworks for HitL should be developed to guide designs that aim to strike a balance between autonomy and control—a satisficing solution, if you will. Our systems analysis of control indicates that each type of control may pose different challenges and require a unique feedback design to balance human control with machine autonomy.

A recent study by OpenAI (Baker et al. 2025) highlighted the challenges in controlling AI meant to detect cases of reward hacking, a phenomenon in which the machine achieves its rewards through behavior that does not align with the designer's intentions. They used chain-of-thought techniques to uncover machine behavior. Alarmingly, not only did the ChatGPT exhibit misaligned behavior, but it also learned to disguise it when prompted to behave as told. Clearly, the technological challenges to enable human control will increase as AI models become more sophisticated and more obfuscated, and humans will have to rely on AI to help control AI (Guan et al. 2024).

Another important shortcoming concerns the feasibility of human control over AI (Greene et al. 2023; Holzinger et al. 2025). Our study demonstrated the feasibility of reciprocal learning to keep the HitL as a prerequisite for control. For the adaptive mode, we offer reciprocal human–machine

learning (RHML) as a method for operationalizing human learning in a human–AI decision system. RHML has been applied to three distinct classification problems: detecting cybersecurity threats (Cohen et al. 2025), identifying illicit drug transactions (Te'eni et al. 2023), and detecting subversive influencers (Lewinski et al. 2024). It can be applied to other forms of complex decision-making in uncertain environments. We are currently expanding our work to support multimodal classification of image and video streams, further enhancing the accuracy of AI threat detection models over time. We are also incorporating generative LLM to improve model explainability and the ease of transferring human knowledge between humans and machines. Ultimately, we aim to design meaningful human control to ensure value alignment by detecting unethical biases in the stable mode of operation and addressing them in an adaptive mode. Moral considerations must evolve in tandem with advancements in AI, ensuring a seamless transition from single-system governance to broader societal integration.

Nevertheless, our approach to gaining human control in human–AI systems has limitations and costs. We began our argument assuming that the systems are designed to serve human goals and preferences. Although widely accepted, this assumption can be contested on the grounds that more intelligent machines may know better what is in the best interest of humans, at least in some situations. Furthermore, keeping HitL has costs that cannot always be justified, and if the primary benefit is gaining control, there may be more effective ways, as noted above. This includes cases where the cost of explainability in terms of reduced accuracy cannot be justified. In addition, human learning may sometimes prove too challenging (Tsamados et al. 2025) or even ineffective compared to advanced machine learning.

Lastly, we build upon the balanced relationship between humans and AI, in which humans neither over-rely on nor dismiss machine feedback. Achieving such a balance is not trivial, as superintelligence becomes increasingly convincing. Feedback design is, thus, crucial in RHML systems. This is the focus of our future research.

5.3 Combining design and regulation to manage risk

The proposed design framework was developed in the context of critical decision-making, where risks are assumed to be high but acceptable. The European AI Act prohibits systems designated as having unacceptable risks, such as those that manipulate behavior (Kieseberg et al. 2023). Our framework addresses control and trusted feedback loops as essential and complementary strategies for managing risk. One enables the other. Alternative strategies have been proposed that overlap with our framework. Kieseberg et al. (2023) proposed several techniques for controllable AI, including providing explainability, sanity checks, using divine (satisficing) rules, and fail-safe procedures that take over control from the autonomous agent.

Antifragility theory (Taleb 2012) is another approach to reducing systemic risk. Antifragility extends beyond Churchman's concept of introducing flexibility in adaptive systems by continually learning in their subsystems using local intelligence, and beyond Ashby's notion of requisite variety; it proactively introduces small failures that create resilience in systems that can benefit from volatility and disorder. This approach must be closely controlled when designing systems that support critical decision-making to avoid disastrous outcomes while training. This may mean introducing dedicated anti-fragility control loops in our design context in each system/subsystem.

Our work has focused on achieving human control by design; however, design should be complemented by regulation (Simon 1996). Regulation is needed to guide and enforce design principles for responsible human–AI interactions; as Article 14 of the European AI Act puts it, "Human oversight shall aim to prevent or minimize the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used …" (EAIA 2025). Beyond compliance, regulation should encourage the co-development of AI governance structures that balance control with adaptability. Transitioning from single-system oversight to broader societal frameworks requires integrating governance policies to ensure that AI remains accountable, timely, and beneficial at scale. Singapore (2025), for instance, has recently expanded its AI governance to include generative AI. Without continual learning in the system, value alignment will fail to consider new values introduced by the country's institutions. Only a coordinated combination of regulation and design can ensure that AI remains both ethically aligned and beneficial to society.

In conclusion, as AI systems increasingly influence critical decision-making, regulatory frameworks must guide their design and deployment to ensure alignment with societal values and governance structures. Establishing clear design methodologies for human–AI collaboration is essential in this context, enabling organizations to define operational boundaries and accountability mechanisms. This will also present an opportunity to delve further into the impact of internal uncertainty due to randomness in the AI models.

Achieving effective, controllable, and trustworthy systems, through design and regulation, is a moving target due to the daily advancements in technology, requiring sound conceptual frameworks and new intelligent technologies. We hope that our framework and the suggested extensions will trigger further investigations into the pressing issue of controlling AI in critical decision-making areas. Moreover, going beyond the challenge of control, AI presents vast opportunities for stimulating continual learning

in human–AI interactions. We built on Churchman's Singerian inquiry systems, which align with Aristotle's notion that learning in general should contribute to a fulfilling life. For AI to enhance human well-being, design methodologies and regulatory structures must evolve to balance efficiency and value-laden oversight, ensuring that AI's role extends beyond optimization to societal enrichment.

Acknowledgements We acknowledge the insightful comments of the editorial team.

Author contributions D.T. wrote the main text, I.Y. wrote parts of it and designed figures, and D.S. wrote and reviewed the text.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Ackoff RL (1974) Redesigning the future: A systems approach to societal problems. John Wiley & Sons Inc, New York

Ågerfalk P. J., Conboy K., Crowston K., Eriksson Lundström J. S., Jarvenpaa S., Ram S., & Mikalef P. (2022) Artificial Intelligence in Information Systems: State of the Art and Research Roadmap. Communications of the Association for Information Systems, 50.

Ashby WR (1956) An introduction to cybernetics. Chapman & Hall, London

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

Baker B., Huizinga J., Gao L., Dou Z., Guan M. Y., Madry A., & Farhi D. (2025) Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv preprint arXiv:2503.11926.

Baskerville RL, Myers MD (2023) Reconceptualizing users: the roles and activities of people as they engage with digital technologies. J Inf Technol 38(4):487–501

Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Darrell T, Mindermann S (2024) Managing extreme AI risks amid rapid progress. Science 384(6698):842–845

Bennetot A, Donadello I, ElQadiElHaouari A, Dragoni M, Frossard T, Wagner B, Diaz-Rodriguez N (2024) A practical tutorial on explainable AI techniques. ACM Comput Surveys 57(2):1–44

von Bertalanffy L. (1968) General Systems Theory: Foundations, Development, Applications. George Braziller NY

Bodor A., Hnida M., & Daoudi N. (2023) Machine Learning Models Monitoring in MLOps Context: Metrics and Tools. International Journal of Interactive Mobile Technologies, 17(23):125-139.

Bommasani R., Hudson D. A., Adeli E., Altman R., Arora S., von Arx S., & Liang P. (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Churchman CW (1968) The Systems Approach. Dell Books, NY

Churchman CW (1971) The design of inquiring systems. Basic Books, NY

Cohen MK, Kolt N, Bengio Y, Hadfield GK, Russell S (2024) Regulating advanced artificial agents. Science 384(6691):36–38

Cohen D, Te'eni D, Yahav I et al (2025) Human–AI Enhancement of Cyber Threat Intelligence. Int J Inf Secur 24:99

Colville I, Pye A, Brown AD (2016) Sensemaking processes and Weickarious learning. Manag Learn 47(1):3–13

EAIA (2025) European AI Act, Article 14. https://artificialintelligenceact.eu/article/14/ (Accessed 5 Feb 2025).

European Union (2025) Shaping Europe's digital future (https://digital-strategy.ec.europa.eu/en) - PDF generated on 03/02/2025.

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

Gal Y., Ghahramani Z. (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In the International Conference on Machine Learning (pp. 1050–1059). PMLR.

Grace K., Stewart H., Sandkühler J. F., Thomas S., Weinstein-Raun B., & Brauner J. (2024) Thousands of AI authors on the future of AI. In arXiv [cs.CY]. arXiv. arXiv:2401.02843.

Greco S., Vacchetti B., Apiletti D., & Cerquitelli T. (2024) Unsupervised concept drift detection from deep learning representations in real-time. IEEE Transactions on Knowledge and Data Engineering, PP(99), 1–14. arXiv:2406.17813.

Greene T, Shmueli G, Ray S (2023) Taking the person seriously: ethically aware IS research in the era of reinforcement learning-based personalization. J Assoc Inf Syst 24(6):1527–1561

Guan M. Y., Joglekar M., Wallace E., Jain S., Barak B., Helyar A., & Glaese A. (2024) Deliberative Alignment: Reasoning enables safer language models. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2412.16339.

Harari Y. N. (2024) Nexus: A brief history of information networks from the Stone Age to AI. Random House, London-New York

Holzinger A., Zatloukal K., & Müller H. (2025) Is human oversight to AI systems still possible?. New Biotechnology, 85, 59-62.

Kieseberg P., Weippl E., Tjoa A. M., Cabitza F., Campagner A., & Holzinger A. (2023) Controllable AI-an alternative to trustworthiness in complex AI systems?. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 1–12). Cham: Springer Nature Switzerland.

Kivimäki J, Nurminen JK, Białek J, Kuberski W (2025) Confidence-based estimators for predictive performance in model monitoring. J Artif Intell Res 82:209–240

Korbak T., Balesni M., Shlegeris B., & Irving G. (2025) How to evaluate control measures for LLM agents? A trajectory from today to superintelligence. In arXiv [cs.AI]. arXiv. arXiv:2504.05259.

Lakshminarayanan B., Pritzel A., & Blundell C. (2017) Simple and scalable predictive uncertainty estimation using deep ensembles.*Advances in neural information processing systems,* Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 1-12.

Lewinsky D, Te'eni D, Yahav-Shenberger IG, Schwartz D, Silverman G, Mann Y (2024) Detecting terrorist influencers using reciprocal human-machine learning: The case of militant Jihadist Da'wa on the Darknet. Humanities Soc Sci Commun 11(1):1–11

Lin J, Ma Z, Gomez R, Nakamura K, He B, Li G (2020) A review on interactive reinforcement learning from human social feedback. IEEE Access 8:120757–120765

London AJ (2019) Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 49(1):15–21

Mikolov T., Chen K., Corrado G., & Dean J. (2013) Efficient estimation of word representations in vector space. In arXiv [cs.CL]. arXiv.http://arxiv.org/abs/1301.3781.

OWASP (2023). OWASP top 10 for large language model applications. The official 1.0.1 release - Full Version. https://owasp.org/www-project-top-10-for-large language-model-applications. Retrieved September 30, 2023. Accessed 06 2025

Pang B., & Lee L. (2004) A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 271–278, Barcelona, Spain.

Reason J (1990) Human error. Cambridge University Press

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

Robbins S (2024) The many meanings of meaningful human control. AI Ethics 4(4):1377–1388

Robbins S. (2025) What machines shouldn't do. AI & Society 40, 4093–4104 (2025). https://doi.org/10.1007/s00146-024-02169-7

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

Russell S (2022) Human-Compatible Artificial Intelligence. Oxford University Press

Santoni de Sio F, Van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robotics AI 5:15

Schoeller F, Miller M, Salomon R, Friston KJ (2021) Trust as extended control: human-machine interactions as active inference. Front Syst Neurosci 15:669810

Shneiderman B (2020) Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Human-Comput Interaction 36(6):495–504

Siebert LC, Lupetti ML, Aizenberg E, Beckers N, Zgonnikov A, Veluwenkamp H, Lagendijk RL (2023) Meaningful human control: actionable properties for AI system development. AI Ethics 3(1):241–255

Silverman et al. (2025). A hybrid mixed methods design of qualitative enhancement and reciprocal feedback loop for augmented text classification. Quality & Quantity. https://doi.org/10.1007/s11135-025-02108-8

Simon H. A. (1996) The Sciences of the Artificial (3rd ed.). MIT Press Cambridge MA

Singapore introduces three new AI governance initiatives. (2025) Retrieved April 2 2025 from https://govinsider.asia/intl-en/article/singapore-introduces-three-new-ai-governance-initiatives

Sparrow R (2024) Friendly AI will still be our master. Or, why we should not want to be the pets of super-intelligent computers. AI Soc 39(5):2439–2444

Suchman L. A. (2007) Human-machine reconfigurations: Plans and situated actions. Cambridge University Press.

Taleb NN (2012) Antifragile: Things That Gain from Disorder. Random House, New York, NY

Te'eni D. (2025) Reciprocal Human–AI Collaboration: Designing Configuration and Delegation for Continual Learning. In The Design of Human-Centered Artificial Intelligence for the Workplace, Coursaris CK, Beringer J, Léger PM, Öz B (eds), Springer, forthcoming.

Te'eni D (1992) Analysis and design of process feedback in information systems: old and new wine in new bottles. Account Manag Inf Technol 2(1):1–18

Te'eni D., Yahav I., Zagalsky A., Schwartz D. G., Silverman G., Cohen D., Mann Y., & Lewinsky D. (2023) Human-Machine Learning: A Theory and an Instantiation for the Case of Message Classification. Management Science. https://doi.org/10.1287/mnsc.2022.03518

Tsamados A., Floridi L., & Taddeo M. (2025) Human control of AI systems: from supervision to teaming. AI and Ethics 5(2)1535–1548 https://doi.org/10.1007/s43681-024-00489-4

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

Wood NG (2024) Explainable AI in the military domain. Ethics Inf Technol 26(2):29

Woods D. D., & Hollnagel E. (2006) Joint Cognitive Systems: Patterns in Cognitive Systems Engineering (1st ed.). CRC Press. https://doi.org/10.1201/9781420005684

Zagalsky A, Te'eni D, Yahav I, Schwartz DG, Silverman G, Cohen D, Lewinsky D (2021) The design of reciprocal learning between human and artificial intelligence. Proceed ACM Human-Comput Interaction 5(CSCW2):1–36