frontiers | Frontiers in Behavioral Economics

TYPE Conceptual Analysis
PUBLISHED 07 September 2023
DOI 10.3389/frbhe.2023.1223281

OPEN ACCESS

EDITED BY
Nicola Lacetera,
University of Toronto, Canada

REVIEWED BY
Bjørn Tallak Bakken,
Inland Norway University of Applied Sciences (INN), Norway
Filippo Pavesi,
University Carlo Cattaneo, Italy

*CORRESPONDENCE
Ivan Đula
✉ ivan.dula@ife.uni-stuttgart.de

†These authors share last authorship

RECEIVED 15 May 2023
ACCEPTED 24 August 2023
PUBLISHED 07 September 2023

CITATION
Đula I, Berberena T, Keplinger K and Wirzberger M (2023) Hooked on artificial agents: a systems thinking perspective. *Front. Behav. Econ.* 2:1223281. doi: 10.3389/frbhe.2023.1223281

Hooked on artificial agents: a systems thinking perspective

Ivan Đula[1,2,3]*, Tabea Berberena[1,2,3], Ksenia Keplinger[4†] and Maria Wirzberger[1,2,3†]

[1]University of Stuttgart, Cluster of Excellence EXC 2075 "Data-Integrated Simulation Science," Stuttgart, Germany, [2]University of Stuttgart, Interchange Forum for Reflecting on Intelligent Systems (IRIS), Stuttgart, Germany, [3]University of Stuttgart, Department of Teaching and Learning with Intelligent Systems, Stuttgart, Germany, [4]Independent Research Group "Organizational Leadership and Diversity," Max Planck Institute for Intelligent Systems, Stuttgart, Germany

Following recent technological developments in the artificial intelligence space, artificial agents are increasingly taking over organizational tasks typically reserved for humans. Studies have shown that humans respond differently to this, with some being appreciative of their advice (algorithm appreciation), others being averse toward them (algorithm aversion), and others still fully relinquishing control to artificial agents without adequate oversight (automation bias). Using systems thinking, we analyze the existing literature on these phenomena and develop a conceptual model that provides an underlying structural explanation for their emergence. In doing so, we create a powerful visual tool that can be used to ground discussions about the impact artificial agents have on organizations and humans within them.

KEYWORDS

artificial agents, artificial intelligence, systems thinking, algorithm appreciation, algorithm aversion, automation bias

1. Introduction

Researchers and practitioners are witnessing the emergence of a new era that integrates the physical world with the digital world and fosters human-machine interactions (Pereira et al., 2023). In this ever-changing world, the pace to produce, work efficiently, and keep up has increased significantly. Since everything seems to be only one click away, the expectation of being better and faster than the competition has been growing steadily. The question is, how does this ever-growing demand companies have to keep up with, affect employees? With current advances in technology and Artificial Intelligence (AI), humans are confronted with Artificial Agents (AAs) in many areas of their lives. While they still have a choice in integrating AAs into their personal and private lives, there is less choice regarding the work environment or public life. Despite AAs being a relatively new technological trend, they are becoming crucial in daily operations of many organizations, especially in manufacturing and service industries (Müller et al., 2021). AAs are implemented in the workplace to complement individual intelligence and thus to increase the quality and accuracy of decision-making processes (Wilkens, 2020). Previous research describes three distinct, though seemingly connected, phenomena among humans interacting with AAs in the workplace: algorithm appreciation, algorithm aversion, and automation bias. Algorithm appreciation occurs when humans appreciate the automated advice and may prefer it to human advice (Madhavan and Wiegmann, 2007; Chugunova and Sele, 2022), even in situations where the automated advice is incorrect and the human advice is correct (Dijkstra, 1999). Algorithm appreciation involves understanding how algorithms work, what they can and cannot do, and how they can be used to solve real-world problems. It also encompasses recognizing the strengths and limitations of different algorithms as well as understanding how to evaluate their performance (Jussupow et al., 2020).

However, not everyone reacts the same (positive) way when it comes to AAs. Some users are growing reluctant to interact with AAs instead of human agents (Jussupow et al., 2020), especially when making complex managerial decisions (Leyer and Schneider, 2019), performing artistic work (Jago, 2019) or selecting new employees (Diab et al., 2011). This phenomenon is known as algorithm aversion (Dietvorst et al., 2018; Castelo et al., 2019; Berger et al., 2021; Chugunova and Sele, 2022). According to Jussupow et al. (2020) algorithm aversion reflects negative behaviors and attitudes toward an algorithm when compared to a human agent and thus assesses an algorithm in a biased manner. Algorithm aversion occurs when people prefer human judgment over algorithmic decision-making, even when algorithms are known to be more accurate and reliable, and can be motivated by a range of factors, such as transparency and explainability (Liao and Fu, 2018).

While algorithm aversion may result in the lack of trust in the automated agent and thus in rejecting its capabilities, there is also empirical evidence that humans may excessively trust AAs, even when the algorithms are proven to be biased or inaccurate (Alberdi et al., 2009). Excessive trust in the AAs means accepting all solutions and suggestions without questioning (Khavas, 2021). This blind trust in algorithms can lead to a phenomenon known as automation bias, where users frequently over-rely on automation, failing to notice errors or discrepancies as well as failing to intervene when they should (Chugunova and Sele, 2022). In contrast to algorithm aversion, one of the reasons for automation bias is a poor understanding of what automation can and cannot do (Itoh, 2010) as well as the attribution of human characteristics and moral judgments to AAs (Ryan, 2020). Human agents experiencing algorithm appreciation use AAs as decision support tools, while those experiencing automation bias use them as a replacement for the human decision-maker. If humans trust algorithms blindly and do not scrutinize algorithmic decisions, automation bias may become a significant issue that leads to harmful outcomes, such as engaging in unethical or self-serving behavior (Chugunova and Sele, 2022) or completely relinquishing decision-making to the AA (Wagner et al., 2018).

Overall, the concepts of algorithm appreciation, algorithm aversion, and automation bias help researchers to understand how humans make decisions when they interact with AAs in organizational settings. Previous research, however, has been rather fragmented focusing on the development of a substantial number of small, niche research topics and offering explanations that cannot be generalized to the whole field. For instance, a design feature of an AA that results in algorithm appreciation in the medical context, may result in algorithm aversion in the context of human resource management. So far, most explanations for the interaction of humans and AAs have focused on the types of tasks for which the AAs are used, design features of AAs, decision authority between human agents and AAs, and human-in-the-loop performance (see Khavas, 2021; Chugunova and Sele, 2022 for a review). This suggests that there is an urgent need in developing a theoretical framework depicting underlying relationships and dynamics of organizational human-machine interactions, which can be used to explain the observed differences and to possibly connect aforementioned fragmented research findings under that framework.

In this paper, we deploy the techniques and tools from systems thinking to develop a theoretical framework that conceptualizes the emergence of algorithm aversion, algorithm appreciation, and automation bias in the context of human-machine interactions in the workplace. The main goal of this paper is to describe the underlying structure that simultaneously enables the phenomena under study to emerge depending on the context, characteristics of artificial agents, and human agents' preferences and biases, therefore providing a better understanding of mechanisms, complex relationships, and dynamics between human agents and AAs in organizational settings. Further, we aim to identify important systemic leverage points that can help us steer the system of human-machine interactions in a desired direction, as well as potentially uncover hidden challenges that might be lurking in the near future as the use of AAs becomes more prominent in organizations.

This research contributes to the existing knowledge in three ways. First, it offers a holistic view over the use of AAs in organizations instead of focusing on "siloed" conceptualizations by addressing social and technical aspects of AI separately. Second, it contributes to the ongoing debate in the human-machine interaction literature on the causes and implications of organizational use of AAs and lays the foundation for future empirical investigations of the effectiveness of different policies for the implementation of AAs in the workplace. Third, this research promotes an interdisciplinary approach for investigating human-machine interactions in organizational settings and sparks new research questions in this area of study.

2. Using systems thinking to understand dynamic systems

Systems thinking can be described as a perspective, a language, and a set of tools, which provide meaningful explanations of the complexity of the world through the lens of wholes and relationships, rather than breaking it down into its parts (Kim, 1999; Sterman, 2000; Ramage and Shipp, 2009). Systems thinking has been applied to a wide range of fields and disciplines due to its ability to solve complex problems, explain non-linear behaviors, understand socio-economic problems, and clarify seemingly illogical behaviors of individuals, countries, and organizations (Monat and Gannon, 2015). A foundational concept within the systems thinking perspective is the so-called *Iceberg Model* (see Kim, 1999), which typically views reality from three different levels of perspective (Figure 1): events, patterns, and systemic structures.

Events are things that happen on a day-to-day basis that we can see or observe. Patterns are sets of consistent and recurring observable events that can reveal recurring trends, when strung together as a series over time. The Iceberg Model argues that events and patterns are caused by underlying and oftentimes hidden, systemic structures. They represent the ways in which the parts of the system are organized. The key aspect about the three different levels of the systems thinking perspective is acknowledging that humans live in an event-oriented world and have an evolutionary predisposition for incorporating events into their mental models of the world, which makes them much easier to notice compared to patterns and systemic structures, and frequently leads to events

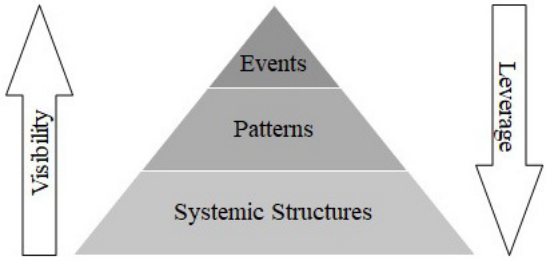

FIGURE 1
The iceberg model.

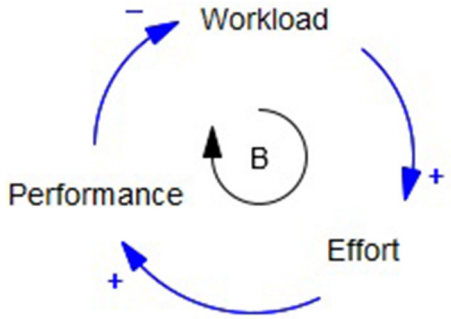

FIGURE 2
A simple workload model.

driving human decision-making (Kim, 1999). However, only when people recognize the underlying structure and its impact on human behavior, then they can truly understand how systems work, what causes them to produce poor results, and how to shift them into better behavior patterns (Meadows, 2015).

One of the first steps in eliciting system structure and attempting to understand system behavior is drawing causal loop diagrams (CLD's). CLD's are one of the most important tools of systems thinking which enable researchers to capture how different system elements (also called variables) are interrelated. They take the form of one or more closed loops that depict cause-and-effect linkages. These loops indicate the presence of reinforcing and/or balancing processes, which determine the behavior of dynamic systems. Reinforcing processes, depicted in a CLD with a reinforcing feedback loop, compound change in one direction with even more change in that same direction. As such, they generate both growth and collapse behaviors. Balancing processes, depicted in a CLD with a balancing feedback loop, seek equilibrium as they endeavor to achieve and maintain a desired state of affairs. Balancing processes generate goal-seeking behavior (Kim, 1999). Figure 2 shows an example of a simple workload model of a human agent within an organization using a CLD. The balancing nature of the feedback loop is highlighted with a letter "B" within a circular-shaped arrow showing the direction in which the variables can be traced around the loop. When there are multiple loops of the same type, they can be differentiated by adding a number identifier to the letter (e.g, B1, B2, R1, R2, etc.).

The model contains three variables: workload, effort, and performance. They are connected to each other with causal links. Workload can be seen as an accumulation of tasks that an individual within an organization is required to process. Within the organizational behavior literature, it can be compared to "job demands", which are aspects of jobs requiring sustained physical, emotional, or cognitive effort (Bakker et al., 2014) and corresponds to the NASA Task Load Index (NASA-TLX)[1] Effort is defined as physical, emotional, or cognitive load that is actually allocated by the human agent to accommodate the demands imposed by the workload. Here we follow Paas et al. (2003) definition of cognitive load, which is seen as a product of task and subject characteristics. Finally, performance can be understood as the rate at which the individual is able to process work-related tasks that directly serve the goals of the organization. In that sense, it is synonymous with "task performance" and "in-role performance" concepts (see Motowidlo and Van Scotter, 1994; Bakker et al., 2012).

The model in Figure 2 can be interpreted as follows: as the workload increases, the effort also increases, hence the plus sign indicating a compounding or positive causal relationship; as the effort increases, the performance increases, once more the plus sign indicating a positive relationship; as the performance increases, the workload decreases, hence the minus sign indicating an opposing or negative relationship; given that an increase in workload results in an opposing reaction, namely a decrease in workload, we are looking at a balancing process, hence the loop symbol with a letter B (see Figure 2). This simple structure provides an explanation for an experience everyone within a given organization goes through. As a human agent is faced with an increased workload, s/he increases the effort. If there is more work to do, individuals tend to be more physically, emotionally, and cognitively involved with their tasks. That has a positive effect on their performance. They are able to process tasks faster, which, given some time, reduces the workload to a desired state. It is important to note that the loop works in the opposite direction as well. If the workload decreases, less effort to handle it is needed, so it also decreases. Decreased effort over some time results in decreased performance, which, given more time, increases the amount of workload as individuals fall behind on some of the work that needs to be done. Furthermore, the causal relationships between variables in this model represent an ideal-world situation. There is plenty of research suggesting that increased effort does not always translate into increased performance. For example, worker burnout is a process in which a hard-working individual becomes increasingly exhausted, frustrated, and unproductive (Homer, 1985). It is possible to extend the model to include these and other dynamics and create a more accurate representation of reality. These dynamics, however, are well understood in the literature and not the focus of our study, which is why we use a simplified understanding of the effort-performance relationship and keep the overall complexity of the model as low as possible. Nevertheless, it should be recognized that they may need to be included in future iterations of the conceptual model to get a more complete picture of the system of human-machine interactions in the workplace,

1 https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf.

particularly if these dynamics are seen as important in a given organizational context.

To illustrate this process, we can use an example of a software developer. Workload here refers to a certain amount of code that the developer is required to generate, translate, explain, and verify, before it can be delivered to a customer. For many reasons, the developer may experience more workload than usual. A new customer may be acquired or a current customer may have new requirements for the existing product, but ultimately it results in an increased workload for the developer. The developer responds by increasing their effort which leads to more lines of code being processed and, ultimately, workload being reduced to the desired level. The key point to note here is that it takes some time between increasing the effort and seeing a decrease in workload. It does not happen instantly but requires commitment and resources.

We use this model as a starting point to then extend it in the following section in order to develop the model of human use of artificial agents in organizations. We rely on recent literature on human-machine interaction in the workplace to elicit the main components of the model, as well as their interconnections. Once completed, we discuss how the same underlying system structure is capable of generating all of the different observed behaviors. We particularly focus on identifying *systems archetypes*. In systems thinking, archetypes are common problem-causing structures that are repeated in many situations, environments, and organizations (Monat and Gannon, 2015). Currently, there are 10 common archetypes and identifying them in dynamic systems is the first step toward changing problematic structures and behaviors (for a detailed overview of systems archetypes see Kim and Anderson, 1998). Our goal is to investigate potential structural origins of automation bias, algorithm aversion, and algorithm appreciation, in order to develop a unifying theoretical foundation for these significantly different behavioral patterns, as well as to identify any underlying systems archetype(s) within the larger system structure, in order to better understand the sources of problematic behaviors and the important leverage points to mitigate them (see e.g., Senge, 1990).

3. Model development

To begin our analysis, we introduce the central variable of our model depicting the use of AAs by individuals in organizational settings (see Figure 3). Similar to other studies investigating implications of human/AA interactions (e.g., Chugunova and Sele, 2022) and in correspondence with the European Commission's AI-Act, we adopt a broad definition of AAs. According to this definition, AAs are advanced systems designed by humans and capable of learning and applying knowledge in new situations (Gams et al., 2019). AAs can, for a given set of human-defined objectives, generate outputs, such as content, predictions, recommendations, or make decisions influencing the environments they interact with (European Commission, 2022). In our model, the use of AA indicates the extent to which a human agent relies on artificial agents in their day-to-day work-related activities. An increase in the use of AA can be interpreted as using artificial agents more often for a particular type of tasks, using artificial agents for more types of tasks, or using more types of artificial agents, depending on the organizational context. Figure 3 extends the simple workload model from the previous section to include the use of AA.

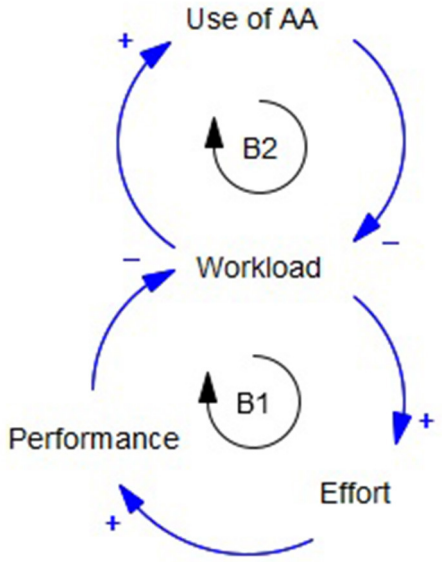

FIGURE 3
The use of artificial agents to reduce the workload.

We propose that there is a positive causal relationship between workload and the use of AA, as well as a negative causal relationship between the use of AA and workload, closing another balancing feedback loop. In particular, if a worker's workload increases above the usual level, they have the ability to manage it through the increase in their effort level; however, this takes time and resources. At the same time, there are new technologies that can help workers achieve their goals more quickly and effectively. One of the main advantages of AAs is that they are able to replace a significant amount of human labor and save time as well as resources. Therefore, if the workload increases, the use of AA also increases, which results in a quick decrease of workload. Apart from advertised benefits of various AA technologies, there is also empirical evidence supporting these causal relationships. For example, Balfe et al. (2015) show that high levels of automation lead to a decrease in subjective workload and the operator activity in the context of rail signaling. To differentiate between the feedback loops, we renamed the original workload loop to B1 and labeled the new loop depicting the use of AA to B2. Both B1 and B2 feedback loops aim to keep the workload of an individual under control and at an acceptable level.

Going back to our software developer example, the B2 loop can be seen as the developer deploying an AA tool, such as GPT-4[2], to take over some of the work-related tasks. As the amount of code that the developer needs to process increases and they want to save time and resources, the developer decides to use GPT-4 to verify that the code works as intended. GPT-4 performs this operation almost instantly, saving the developer a lot of time, keeping their

2 https://openai.com.

workload at the usual level, and eliminating the need to drastically increase their effort level.

In a perfect scenario, this is where our analysis would end. When facing an increased workload, human agents would respond by increasing their effort and, if necessary, by deploying AAs in order to process their tasks and achieve a desired level of workload. However, previous research suggests that there are side effects of the use of AAs. One of the most consistent and impactful effects is the reduction in emotional and social response of human agents interacting with AAs. Chugunova and Sele (2022) summarize the abundant empirical evidence for this effect as stemming from both subjective and behavioral measurements, typically manifested in less immediate emotional reaction to the AAs actions, as well as generally decreased levels of emotional arousal in human-machine interactions (Teubner et al., 2015). Simply said, human agents respond differently to AAs than to other human agents; primarily they show less emotional and social engagement, which can potentially have serious negative consequences.

Melo et al. (2016) suggest that human agents feel considerably less guilt when interacting with an AA, which makes them more willing to exploit machines in an unethical or illegal manner. Moore et al. (2012) show that morally disengaged human agents in organizations are increasingly more likely to engage in various unethical behaviors. Self-reported unethical behavior, fraud, self-serving decisions in the workplace, supervisor- and coworker-reported unethical behavior, Machiavellianism or the idea that the end justifies the means (Jones and Paulhus, 2009), and other similar types of behaviors are more common when the engagement is low. Köbis et al. (2021) further highlight that AA acting in an influencer or an advisor role have many characteristics that enable human agents to reap the benefits of unethical or illegal behavior, while still feeling good about themselves. Finally, Corgnet et al. (2019) indicate that when exposed to a working environment with AAs instead of other human agents, humans are more likely to behave in a manner that reduces their overall effort in a workplace and compromises their performance. We now add these relationships to our model as shown in Figure 4 while noting once again that research is not necessarily unified on all of them.

We can observe that the model contains a new feedback loop. As the use of AA increases, emotional and social response of human agents decreases. This is likely to increase undesirable behavior, which leads to a decrease in effort and subsequently performance (see Figure 4). Reduced performance means that, over time, a worker processes their work tasks slower and the quality of their work declines; therefore, their workload increases. Increased workload leads to an increase in the use of AA. This means that an increase in the use of AA results in further increase in the use of AA, which indicates a reinforcing feedback loop (R1).

Going back to our software developer example, the R1 loop represents a shift toward problematic work behavior, which ultimately makes the developer more dependent on the use of AA. For example, the developer is someone who genuinely enjoys their work. They like analyzing their customer's problems, thinking about solutions, and implementing those solutions in the form of a code. New and exciting problems are what keeps the developer emotionally and socially engaged with the work and are the main reason to engage in beneficial working behaviors. This could

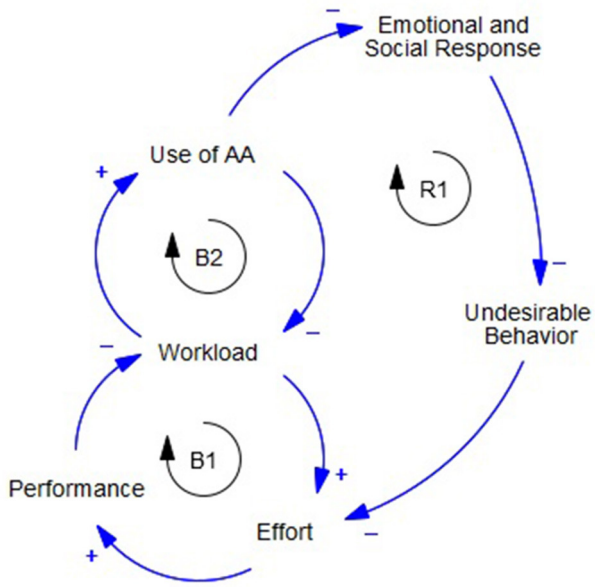

FIGURE 4
Exploitation of artificial agents.

mean that they double check the coding, visit support forums for developers, discuss issues with other colleagues, frequently reach out to the customer for feedback, try out multiple solutions to see which one is best, and so on. All of these activities keep the effort level high and result in higher quality of work. Deploying GPT-4 to take over some or the coding activities, over some time, can change how the developer interacts with their work. Because GPT-4 offers quick fixes, the developer has less exposure to unique challenges and creative solutions that make the work enjoyable. Instead of double-checking the work or discussing problems with colleagues and customers, the developer tends to accept the solutions suggested by GPT-4 without reservation. Because of that, the amount of effort exerted by the developer decreases. They no longer regularly visit support forums or try out multiple solutions as it is no longer necessary. This new lower level of effort becomes the norm and the performance decreases as a result, workload increases, and, finally, the developer needs to use GPT-4 even more to compensate for that. This starts a vicious cycle where the developer relies more and more on GPT-4 to do the coding work.

Another consequence of reduced emotional and social response is an increase in human agent's rationality of interactions. As Chugunova and Sele (2022) note, it has been shown that the introduction of AAs helps human decision makers make more rational decisions, ultimately leading to economic benefits. We model this relationship by adding the rationality of interactions variable to the model. This adds two new causal links: a negative link between emotional and social response and rationality of interactions as well as a positive link between rationality of interactions and performance. As shown in Figure 5, this creates a new balancing feedback loop (B3).

10.3389/frbhe.2023.1223281

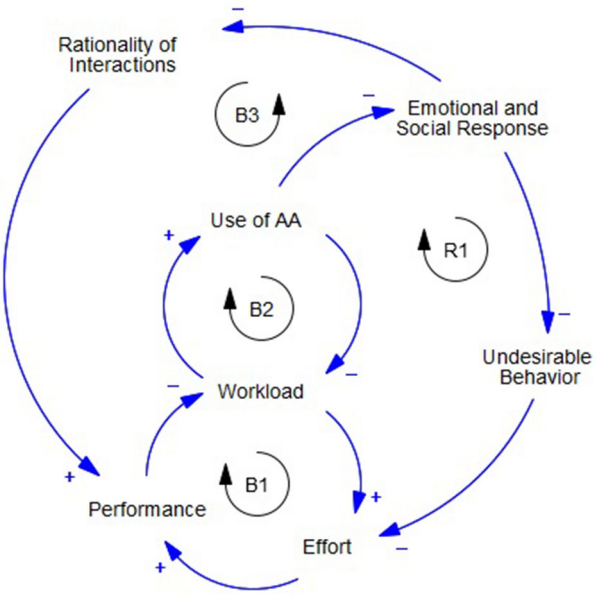

FIGURE 5
Boosting performance through artificial agents.

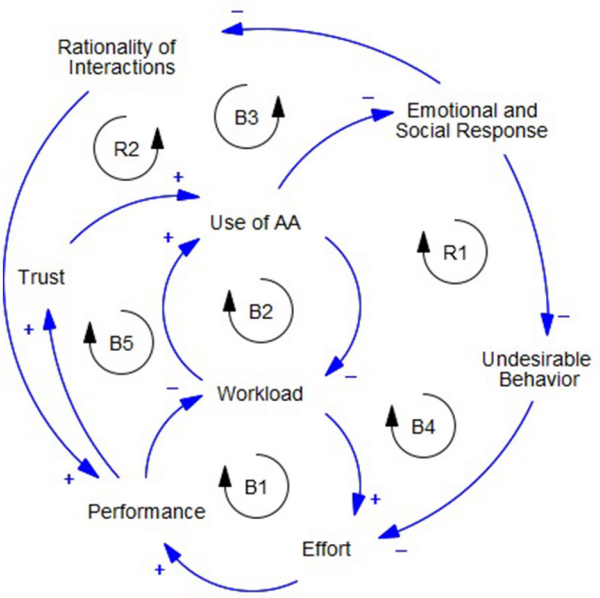

FIGURE 6
The model of use of artificial agents in organizations.

In the software developer example, the B3 loop represents tangible benefits of using GPT-4. Using the AA helps the developer write a better code, as the tool is quickly drawing solutions from huge amounts of existing examples. It makes the developer more efficient in their work-related tasks, boosting their performance and reducing the workload. Reduced workload further reduces the need to use GPT-4.

As stated earlier, the use of AA is the central variable of our model and is primarily driven by the workload variable which establishes a potential need for AA. However, literature in this field has highlighted another important driver of the use of AA, namely trust. Trust in AI technologies is seen as a critical component of successful integration of AA into organizations (Glikson and Woolley, 2020). It has been identified as one of the main influencers of automation use (Parasuraman and Riley, 1997). Low trust in AAs and the subsequent refusal of humans to use AAs in the workplace can be predominantly explained by the poor performance of AAs in given settings (see e.g., Dzindolet et al., 2003; Dietvorst et al., 2015; 2018; Dietvorst and Bharti, 2019). In other words, if the use of AA results in poor performance, humans tend to lose trust in the AA's ability and use them less. This indicates positive causal relationships between trust and the use of AA, as well as performance and trust. We therefore extend our model to include these relationships as seen in Figure 6.

Adding trust to the model creates three new feedback loops: two balancing loops and one reinforcing loop. Both B4 and B5 loops further amplify previously described mechanisms through which effort is reduced, leading to a decrease in performance. This results in a loss of trust in AA, and ultimately to the refusal to use them in the workplace. The B4 loop describes how effort decreases through reduced emotional and social response. As the use of AA increases, emotional and social response of human agents decreases, while their undesirable behavior increases, resulting in a decrease in effort. The B5 loop describes reduced effort through workload (see Figure 6). As the use of AA increases, workload of a human agent decreases, meaning they can reduce their effort level. The reinforcing loop (R2) captures the performance boost obtained through the use of artificial agents. As the use of AA increases, emotional and social response of human agents decreases, while rationality of interactions increases and results in performance increase. Increased performance leads to an increase in trust, and ultimately an increase in the use of AA. A summary of causal pathways can be found in Table 1.

TABLE 1 Summary of causal pathways.

Loop	Causal relationships
B1	Workload→ (+)Effort→ (+)Performance→ (-)Workload
B2	Workload→ (+)Use of AA→ (-)Workload
B3	Workload→ (+)Use of AA→ (-)Emotional and Social Response→ (-)Rationality of Interactions→ (+)Performance→ (-)Workload
B4	Use of AA→ (-)Emotional and Social Response→ (-)Undesirable Behavior→ (-) Effort→ (+)Performance→ (+)Trust→ (+)Use of AA
B5	Workload→ (+)Effort→ (+)Performance→ (+)Trust→ (+)Use of AA→ (-)Workload
R1	Workload→ (+)Use of AA→ (-)Emotional and Social Response→ (-)Undesirable Behavior→ (-)Effort→ (+)Performance→ (-)Workload
R2	Use of AA→ (-)Emotional and Social Response→ (-)Rationality of Interactions→ (+)Performance→ (+)Trust→ (+)Use of AA

Applying this to the case of the software developer, the three loops of B4, B5, and R2 explain a relationship between a human agent and their trust toward GPT-4. While the B4 and B5 loops discourage the use of GPT-4 due to reduced trust in the tool, the R1 loop encourages the use of GPT-4 due to the increase in workload (see Figure 6). For example, as the developer increases the use of GPT-4, they may discover more errors in the code written by GPT-4, which increases the amount of work to be re-done and reduces the level of trust in the tool. On the other hand, the R2 loop may counter the balancing effect of the B3 loop. The B3 loop discourages the further use of GPT-4 because the developer has already been able to significantly reduce the workload due to GPT-4's effectiveness. The R2 loop, however, increases the developer's willingness to use GPT-4 as a direct consequence of increased trust in this tool and its ability to boost performance. For example, after deploying GPT-4 and seeing how effective it was in verifying the code, the developer's trust in the tool increases and they become more likely to use it for other purposes besides verification.

It is important to note that this version of the model does not capture the full extent of the real-world system. There are more variables and causal links, as well as external influences, that can and should be added to it if we wish to incorporate all of the relationships established in the literature. We discuss some of these potential extensions in the concluding section. In its current form, however, the model accomplishes our main objectives. It provides a structural foundation for explaining often conflicting observations when it comes to the use of AA in organizational settings and enables us to detect archetypal structures that can generate problematic behavior within organizations. Finally, it allows us to identify critical points in the system that can be used as leverage points to steer its behavior in a desired direction. In the upcoming section, we will provide a more detailed explanation and expansion of each of these points.

4. Dynamics of human-machine interaction

As mentioned before, a foundational concept of system thinking is that the underlying structure drives the behavior of a dynamic system. The model described in the previous section provides a visualization of the organizational structure for the implementation of AAs. This naturally raises the question what are the potential behaviors this structure can generate. While explaining the structure of the model and using the example of the software developer, it became apparent that the response to this question is not a simple one, nor is it straightforward. It likely depends on a multitude of unique characteristics of each individual system. Some of the loops we identified work to compound change in the system, while others oppose it. Some links amplify behaviors generated by other loops. The strength of individual loops, meaning how quick they work and how effectively they are able to change the central variables in the system, determines in which direction the system will develop. One part of the structure, however, corresponds to a typical problem-generating structure, commonly observed in dynamic systems. The B1, B2, and R1 loops together form a systems archetype known as "shifting the burden", or "addiction" when it creates particularly perilous consequences. In Figure 7, we show the generic structure of the archetype.

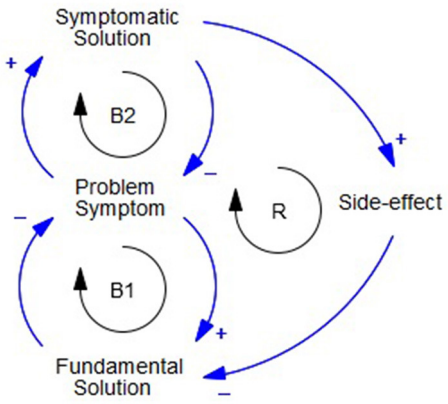

FIGURE 7
Generic structure of shifting the burden/addiction archetype.

Shifting the burden usually starts with a problem symptom that prompts us to intervene and solve it (Kim and Anderson, 1998). In our model, this role is filled by workload. Specifically, a higher than desired workload is the problem an individual aims to solve. To address this, we apply a symptomatic solution that indeed eases the problem symptom for a while (see B2 loop in Figure 7). Individuals increase their use of AA to reduce the workload. After human agents apply the symptomatic solution and the problem symptom decreases, they may feel no need to adopt a more difficult, time-consuming fundamental solution (see B1 loop in Figure 7). Individuals do not need to increase their effort level in order to improve their performance (a fundamental solution), as their workload is quickly managed through the use of AA. The symptomatic solution, however, also has a negative side-effect that contributes to the decrease of the ability to implement a fundamental solution (see R loop in Figure 7). Although the fundamental solution requires more time and effort, it is more likely to solve the problem at the root-cause level and prevent it from recurring. Each application of the symptomatic solution decreases the ability of individuals to implement a fundamental solution through a reinforcing process (Kim and Anderson, 1998). Over the course of time, as the problem symptom increases, the reliance of human agents on a symptomatic solution increases, while their ability and willingness to deploy the fundamental solution declines (see Figure 8).

In our model, the problem symptom is the increasing workload, while the use of AA is the symptomatic solution for the problem symptom. The negative side-effect is the decrease in emotional and social response of human agents, which leads to an increase in various undesirable behaviors, ranging from unethical to illegal ones (Moore et al., 2012). The more a given individual relies on AA to reduce their workload, the less emotionally and socially engaged they become, and the more undesirable working behaviors they adopt. Ultimately, this further reduces their effort level, making the application of the fundamental solution more difficult. Kim and Anderson (1998) suggest that the best way to manage a "shifting the burden" situation is to avoid it completely or to prevent it

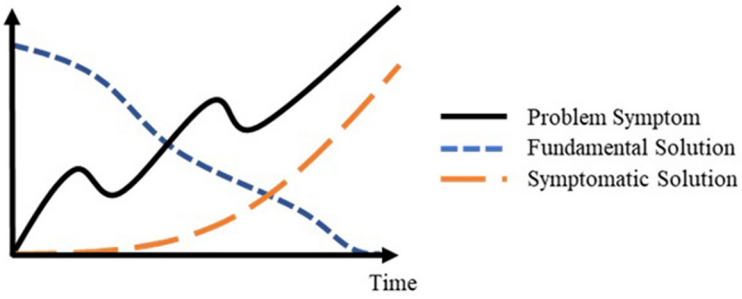

FIGURE 8
Behavior over time of symptomatic and fundamental solutions.

from becoming entrenched. Human agents should pay attention to the pressures that push them into responding automatically rather than thoughtfully and notice when they are responding primarily to relieve pressure rather than to address a problem. If the same problems seem to be recurring over and over, despite the attempts to solve them, then people might be in the "shifting the burden" situation and should look for the deeper causes of the problems, as well as potential side-effects of the proposed solutions.

Remarkably, as shown in Figure 6, there are also some benefits of the reduced emotional and social response of human agents, which create new feedback loops and may help mitigate some of the negative side-effects of the "shifting the burden" structure. AAs have the ability to improve productivity of human agents significantly, so that they can actually reduce the need to use AA through the B3 loop. Working alongside AA can make human decisions more rational and improve performance without having to increase effort. Therefore, even if the effort level and the quality of work decrease because of the adoption of the symptomatic solution (e.g., use of AA), the performance benefits from increased rationality of interactions might be enough to bring the workload to an acceptable level and limit the further use of AA. For this reason, it is important to consider how strong of an effect each feedback loop has in a given context. If the AA provides fast solutions of high quality for solving the initial problem, while simultaneously discouraging continuous reliance on it, humans can avoid the negative side-effect of adopting undesirable behaviors and keep their effort and performance at acceptable levels. If the AA encourages humans to constantly rely on itself, then the addiction cycle may start and humans may become dependent on and less critical of the use of that particular AA.

The overall behavior of the human-machine interaction system greatly depends on the strength of the two side-effect loops, going through the variables of rationality of interactions and undesirable behavior. Depending on the context, the impact of the characteristics of artificial agents as well as the impact of an individual's own preferences and biases (Kahneman, 2011) on the strength of loops can range from weak, meaning it does not have a significant impact, to strong, meaning it does have a significant impact. Figure 9 suggests four potential scenarios describing how the strength of the two side-effect loops can influence the system response. Initially, the amount of workload is constant at a value that can be considered normal. The human agent applies consistent effort appropriate for that level of workload and the use of AA is negligible. We consider a situation in which, due to external influences, the amount of workload increases significantly at a given point in time. Using the model in Figure 6, we can make predictions about the likely progression of workload, effort, and use of AA variables for each scenario.

In the first scenario, we consider a situation where both loops (rationality of interactions and undesirable behavior) have a strong effect on the behavior of the system (upper right corner in Figure 9). After the increase in workload, the use of AA increases. Initially, the effort level stays about the same as before. The workload has increased but the use of AA has a strong positive effect on performance, eliminating the need to significantly increase the effort level. However, the strong effect on undesirable behavior means that, given some time, the effort level of the human agent begins to decrease. As the workload reaches normal levels again, the use of AA and effort stabilize as well, with the use of AA remaining above the initial level and the effort below. Simply put, using the AA has solved the original problem, namely the increased workload, however, it has created opportunities for the human agent to engage in undesirable behaviors and a desire to reduce the overall effort level. Because the rationality of interactions has a strong effect on the system, it allows for this reduction and AA permanently takes over a significant portion of the workload. This situation, in which the human agent relies on AAs more than necessary and puts in less than appropriate amount of effort, corresponds to observations related to automation bias. As mentioned before, automation bias typically manifests itself "as a heuristic replacement for vigilant information seeking and processing" (Mosier et al., 1996, p. 205). In other words, individuals fail to exert the appropriate level of effort to process their tasks and instead over-rely on artificial agents.

In the second scenario, we consider a situation where the rationality of interactions loop has a strong effect and the undesirable behavior loop has a weak effect on the behavior of the system (upper left corner in Figure 9). Similar to the previous scenario, the increase in workload is followed by an increase in the use of AA. Once again, the effort does not change much as the increased workload and highly effective artificial agents cancel each other out. This time, however, there is no significant negative effect when it comes to effort. Because the undesirable effect loop is weak, the human agent only experiences a slight decrease in effort, which is the result of a significant boost in the human agent's

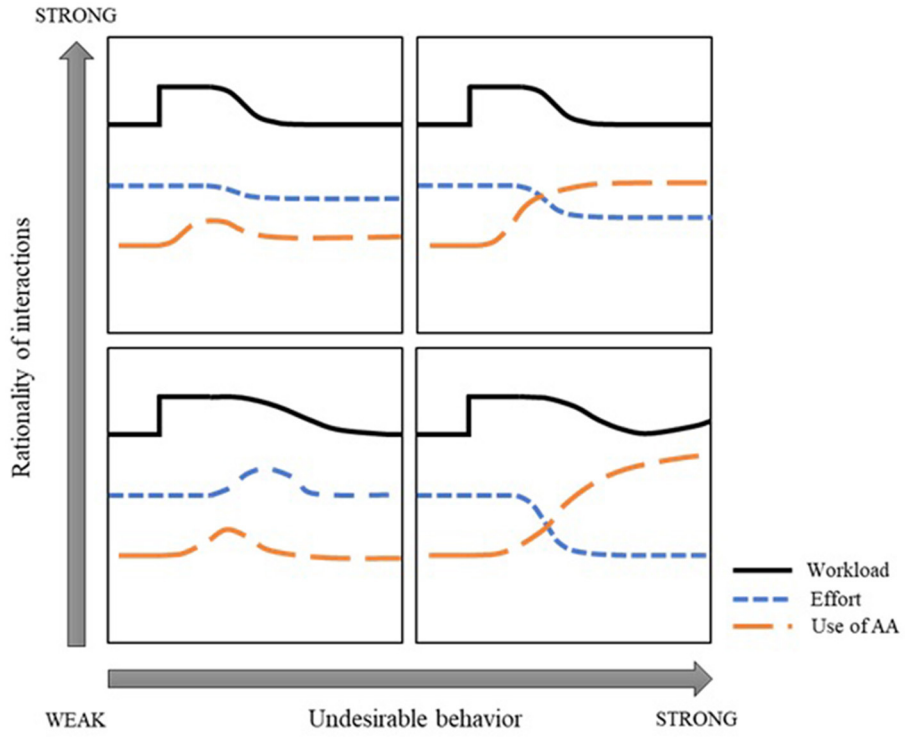

FIGURE 9
Use of artificial agents as a product of feedback loops' strength.

performance. As soon as the workload reaches the desired level, the use of AA decreases, but it stays above the initial level due to increased trust in AA. The effort level, therefore, will decrease slightly in the long run, but will stay high enough for the human agent to exert an appropriate level of oversight and not over-rely on AA. In many cases, this would be considered an ideal scenario for the use of human-machine interfaces in the workplace and precisely what many of the proponents of AI promise. It corresponds to the concept of algorithm appreciation (Madhavan and Wiegmann, 2007) when people exhibit preference for the use of AA without allowing them to fully replace their own decision making. Our model suggests that, in order to achieve this objective, the benefits of using AA should be quickly observable, while potential avenues for the development of undesirable behaviors should be closely monitored and discouraged.

In the third scenario, we consider a situation where both loops have a weak effect on the behavior of the system (bottom left corner in Figure 9). The increase in workload is again followed by an increase in the use of AA. However, because the artificial agent does not improve a human agent's performance significantly, there is an increase in distrust toward it, resulting in reluctance to continue using it. Instead, the human agent increases their own effort to solve the workload issue. It takes longer than with the use of AA, but eventually the workload is reduced to the desired level and the use of AA drops down to the initial level. The entire system returns to the initial state with increased distrust toward AA. This situation corresponds to the concept of algorithm aversion (Jussupow et al., 2020), where human agents oppose the use of AA, even when there are some benefits or when there are no significant downsides to their use. Our model suggests that this situation occurs when the observed benefits of AA are not applicable to the users or when they are not available fast enough.

Finally, in the fourth scenario, we consider arguably the worst situation imaginable, where the rationality of interactions loop has a weak effect and the undesirable behavior loop has a strong effect on the behavior of the system (bottom right corner in Figure 9). In this situation, once the workload and the use of AA increase, there is a strong negative effect on the effort due to the increase in undesirable behavior. Even though AAs are not particularly effective, the human agent continues to increase their use to compensate for the declining level of effort. Initially, this approach works and the workload of the human agent decreases. However, relying more on the ineffective AA begins to create new problems. The human agent no longer exerts adequate oversight on the AA and allows it to replace a portion of their decision making. The likelihood of lower quality work and mistakes increases, resulting in more rework. The workload begins to increase again forcing the human agent to rely even more on AA. The resulting behavior can be best described through shifting the burden archetype where the human agent becomes addicted to the use of AA.

Unlike the other three scenarios where automation bias, algorithm appreciation, and algorithm aversion could be connected to model structure, there is no corresponding situation identified in the AA literature that identifies addiction to the use of AA. This may be simply due to the fact that AA technologies are still new and at the beginning of their diffusion into organizations.

It is possible that none of the AA technologies simultaneously create significant opportunities for undesirable behavior while not significantly improving performance. Our conceptual model illustrates the system structure allows for this type of behavior and organizations should be careful when introducing new AA technologies. This is precisely what the European Commission's AI-Act (European Commission, 2022) encourages when it advocates for considering the level of risk new technologies bring about to be able to counteract it if needed. Even if there are no known AA technologies that function like in the fourth scenario, given their rapid recent development, they are bound to appear sooner or later. It is difficult to consider a positive outcome in a situation where the benefits of AA are weak, while the opportunities to engage in undesirable behaviors are plentiful.

5. Discussion

As AAs permeate organizations, they inevitably come into contact with human employees and change the way operations are conducted. Researchers have found that the human response to AAs typically falls into one of three categories: 1. human agents reject and distrust AAs and prefer advice from other humans even when AAs' performance is superior (algorithm aversion); 2. human agents accept and prefer AAs' advice over human advice even when AAs' performance is inferior (algorithm appreciation), and 3. human agents allow AAs to completely take over decision-making without exerting appropriate oversight (automation bias). A significant progress has been made to explain the circumstances under which these behaviors emerge. Nevertheless, there is still a lack of understanding of the underlying mechanisms on a systemic level that serve as a source for the observed behaviors. Employing systems thinking to elicit the underlying system structure allows us to develop a fundamental understanding of the causes for the emergence of algorithm appreciation, algorithm aversion, and automation bias in different contexts as well as to uncover mechanisms through which the system of organizational human-machine interactions can be governed more effectively.

In this paper, we analyze the existing literature on human-machine interactions in the workplace in order to develop a conceptual model of the use of AAs in organizations. We identify a number of variables and possible causal relationships that, once connected, form different balancing and reinforcing feedback loops that govern the system behavior. Previous research has consistently identified a decrease in emotional and social response as a consequence of the increased use of AAs. In its turn, this emotional decrease may lead to an increase in undesirable behavior (negative outcome) and, at the same time, may increase rationality of interactions (positive outcome). We suggest that these effects are parts of two separate feedback loops, one reinforcing and one balancing, that affect the use of AAs. The undesirable behavior loop generates pressure toward further increasing the use of AAs, while the rationality of interactions loop counteracts the increase in the usage. Furthermore, the undesirable behavior loop forms a specific systems archetype called shifting the burden. This common structure typically generates addictive behaviors where human decision-makers become increasingly dependent on quick interventions into systems in order to resolve issues, rather than opting for resource intensive and time-consuming fundamental solutions. In addition to these two feedback loops and the archetype, there are further balancing and reinforcing feedback loops that include trust as a core component influencing the use of AAs in organizations.

The newly developed conceptual model is capable of explaining the emergence of algorithm aversion, algorithm appreciation, and automation bias through the interplay of its feedback loops. If both the undesirable behavior and the rationality of interactions loop exert strong influence on the system by simultaneously providing significant performance increase and opportunities for undesirable behavior, system dynamics will likely be inclined toward the automation bias. If the rationality of interactions loop exerts strong influence, and the undesirable behavior loop exerts weak influence on the system dynamics, it is expected to lean toward algorithm appreciation. Finally, if both feedback loops exert weak influence on system dynamics, it is expected to observe a tendency toward algorithm aversion. Through our model, we suggest that the observed behaviors related to human-machine interactions in the workplace are a matter of compounding and/or suppressing these two feedback loops. For example, previous research suggests that humans are less averse toward the use of AAs when they provide humans with a recommendation rather than a decision (Bigman and Gray, 2018). Our model sheds light on why this happens. Requiring input from a human decision-maker has a positive influence on effort, which significantly weakens the influence that the undesirable behavior loop exerts on the system. At the same time, it has a positive effect on trust, which encourages the use of AAs and increases rationality of interactions through the emotional and social response, strengthening the influence of the rationality of interactions loop. Therefore, by requiring more human intervention, the system shifts toward the algorithm appreciation quadrant, as depicted in Figure 9.

Similarly, our model can serve as a visual tool fostering the discussion on other findings and factors influencing the use of AAs, such as responsibility, context, expectations (Chugunova and Sele, 2022), type of task (Khavas, 2021), and perceived control (Green and Chen, 2019). It also highlights the importance of considering other aspects of AA technologies besides their impact on performance. As our model shows, even if human agents use a high performing tool, the overall system behavior may be problematic. Factors mitigating emotional and social response, undesirable behavior, effort, and trust are just as important as the quality of technology being used. From the organizational perspective, however, these factors might be more manageable and represent better leverage points in the system. An interesting topic to consider is related to AAs' design, specifically anthropomorphism or making the physical appearance and characteristics of AAs more human-like. It has been known for some time that as machines are made to appear more human-like, human agents' emotional response to them becomes increasingly positive and empathetic, however, there is a point beyond which this turns into strong revulsion (Mathur and Reichling, 2016). As the appearance continues to become less distinguishable from a human, the emotional response becomes positive again and approaches human-to-human empathy levels. This phenomenon known as the "Uncanny valley" (Mori, 2012) can have significant implications when it comes to understanding

the relationship between the use of AAs and emotional and social response.

The relevance and importance of our conceptual model can be further highlighted by the European Commission's efforts to regulate AI technologies, particularly regarding high-risk technologies as defined in the AI Act (European Commission, 2022). For example, it identifies systems that use AI technologies in hiring, human resource management, and access to self-employment as high-risk systems and subjects these technologies to strict obligations before they can be put on the market. Most notable obligations include ensuring appropriate human oversight measures to minimize risk and providing clear and adequate information to the user. As our model indicates, measures like these can certainly have a positive effect on the system as they support the emergence of algorithm appreciation. What should not be ignored by the regulators however, is the need to simultaneously suppress the opportunities for undesirable behaviors as that could lead toward automation bias or addiction to AAs. Moreover, some consideration should be given to the development of similar regulations or guidelines for the implementation of AA technologies targeting the organizational level. The model we presented here can serve as a useful tool in that effort.

To develop the conceptual model of human-machine interactions in the workplace, we build on empirical evidence and shared knowledge about the factors contributing to algorithm aversion, appreciation and automation bias. This approach allows us to evaluate the mechanisms through which different factors shape the system behavior. While considering the most relevant endogenous variables, it is important to note that there are other external variables that might also play a role in the usage and acceptance of AAs, such as perceived control (Green and Chen, 2019), judgment (Chugunova and Sele, 2022) as well as explainability (Andras et al., 2018; Samek et al., 2019). For the purpose of this research, however, we decided to focus on a small set of relevant variables as the inclusion of further endogenous variables would be beyond the scope of the current conceptual model. Not only would that make the model less understandable, but it would also prevent us from making any conclusions regarding expected progression of focus variables. Future research could benefit from investigating the interplay between new endogenous variables and the core model to better understand the full extent of AAs' impact on humans in organizations. Future research can also use our conceptual framework to develop a quantitative simulation model. This would help researchers to better understand the relationships between the system structure and the observed behaviors, identify most impactful leverage points, and test different policies aimed at maximizing benefits of the use of AAs in organizations, while simultaneously minimizing the accompanying risks.

Author contributions

IĐ: Conceptualization, Resources, Methodology, Visualization, Writing—original draft, and Writing—review and editing. TB: Conceptualization, Resources, Writing—original draft, and Writing—review and editing. KK: Conceptualization, Supervision, Project administration, and Writing—review and editing. MW: Conceptualization, Funding acquisition, Resources, Supervision, and Writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2075–390740016, and the reference number UP 31/1 for the Stuttgart Research Focus Interchange Forum for Reflecting on Intelligent Systems (IRIS).

Acknowledgments

We acknowledge the support by the Stuttgart Center for Simulation Science (SC SimTech).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alberdi, E., Strigini, L., Povyakalo, A. A., and Ayton, P. (2009). "Why are people's decisions sometimes worse with computer support?" in *SAFECOMP 2009, LNCS 5775*, eds. B. Buth, G. Rabe, T. Seyfarth (Heidelberg: Springer Verlag Berlin), 18–31. doi: 10.1007/978-3-642-04468-7_3

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., et al. (2018). Trusting intelligent machines: deepening trust within socio-technical systems. *IEEE Technol. Soc. Magazine* 37, 76–83. doi: 10.1109/MTS.2018.2876107

Bakker, A. B., Demerouti, E., and Sanz Vergel, A. I. (2014). Burnout and work engagement: the JD-R approach. *Ann. Rev. Org. Psychol. Org. Behav.* 3, 389–411. doi: 10.1146/annurev-orgpsych-031413-091235

Bakker, A. B., Tims, M., and Derks, D. (2012). Proactive personality and job performance: the role of job crafting and work engagement. *Human Relat.* 10, 1359–1378. doi: 10.1177/0018726712453471

Balfe, N., Sharples, S., and Wilson, J. R. (2015). Impact of automation: Measurement of performance, workload and behaviour in a complex control environment. *App. Ergon*. 47, 52–64. doi: 10.1016/j.apergo.2014.08.002

Berger, B., Adam, M., Rühr, A., and Benlian, A. (2021). Watch me improve–Algorithm aversion and demonstrating the ability to learn. *Bus. Inform. Sys. Engineering* 63, 55–68. doi: 10.1007/s12599-020-00678-5

Bigman, Y. E., and Gray, K. (2018). People are averse to machines making moral decisions. *Cognition* 181, 21–34. doi: 10.1016/j.cognition.2018.08.003

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependant algorithm aversion. *JJ. Marketing Res*. 56, 809–825. doi: 10.1177/0022243719851788

Chugunova, M., and Sele, D. (2022). We and it: an interdisciplinary review of the experimental evidence on how humans interact with machines. *J. Behav. Exp. Econ*. 99, 1897. doi: 10.1016/j.socec.2022.101897

Corgnet, B., Hernán-Gonzalez, R., and Mateo, R. (2019). *Rac(g)e Against The* Machine? *Social Incentives When Humans Meet Robots*. GATE WP.

Diab, D. L., Pui, S. Y., Yankelevich, M., and Highhouse, S. (2011). Lay perceptions of selection decision aids in U.S. and non-U.S. samples. *Int. J. Select. Assess*. 19, 209–216. doi: 10.1111/j.1468-2389.2011.00548.x

Dietvorst, B., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. journal of experimental psychology. *General* 144, 114–126. doi: 10.1037/xge0000033

Dietvorst, B. J., and Bharti, S. (2019). *Risk Seeking Preferences Lead Consumers to Reject Algorithms in Uncertain Domains*. ACR North American Advances.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Manag. Sci*. 64, 1155–1170. doi: 10.1287/mnsc.2016.2643

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behav. Inform. Technol.* 18, 399–411. doi: 10.1080/014492999118832

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Human-Comp. Studies*. 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7

European Commission (2022). Ethics guidelines for trustworthy AI. Proposal of regulation 2021/0106. Available online at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed April 17, 2023).

Gams, M., Gu, I. Y., Härmä, A., Muñoz, A., and Tam, V. (2019). Artificial intelligence and ambient intelligence. *J. Ambient Intell. Smart Environ*. 11, 71–86. doi: 10.3233/AIS-180508

Glikson, E., and Woolley, A. (2020). Human trust in artificial intelligence: review of empirical research. *Acad. Manag. Annals*. 3, 1–91. doi: 10.5465/annals.2018.0057

Green, B., and Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceed. ACM Human-Comp. Int*. 3, 1–24. doi: 10.1145/3359152

Homer, J. B. (1985). Worker burnout: a dynamic model with implications for prevention and control. *Sys. Dyn. Rev*. 1, 42–62. doi: 10.1002/sdr.4260010105

Itoh, M. (2010). Necessity of supporting situation awareness to prevent over-trust in automation. *Int. Elect. J. Nucl. Safety Simulat*. 2, 150–157.

Jago, A. S. (2019). Algorithms and authenticity. *Acad. Manag. Discov*. 1, 38–56. doi: 10.5465/amd.2017.0002

Jones, D. N., and Paulhus, D. L. (2009). "Machiavellianism," in Handbook of Individual Differences in Social Behavior, eds M. R. Leary and R. H. Hoyle (The Guilford Press), 93–108.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. Research papers. 168. Available online at: https://aisel.aisnet.org/ecis2020-rp/168/

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Khavas, Z. R. (2021). A review on trust in human-robot. *Interaction*. 4, 45. doi: 10.48550./arXiv.2105.10045

Kim, D. H. (1999). Introduction to Systems Thinking. Pegasus Communications, Inc., Waltham, Massachusetts. Available online at: www.pegasuscom.com

Kim, D. H., and Anderson, V. (1998). *Systems Archetype Basics: From Story to Structure* Waltham, Massachusetts: Pegasus Communications, Inc.,.

Köbis, N., Bonnefon, J. F., and Rahwan, I. (2021). Bad machines corrupt good morals. *Nat. Hum. Behav*. 5, 679–685. doi: 10.1038/s41562-021-01128-2

Leyer, M., and Schneider, S. (2019). *Me, You or Ai? How Do We Feel About Delegation. Proceedings of the 27th European Conference on Information Systems (ECIS)*. doi: 10.5465/AMBPP.2019.13580abstract

Liao, Q. V., and Fu, W. T. (2018). Do people trust algorithms more than companies realize? Proceedings of the *2018 CHI Conference on Human Factors in Computing Systems*, 447.

Madhavan, P., and Wiegmann, D. A. (2007). Similarities and differences between human– human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*. 8:4, 277–301. doi: 10.1080/14639220500337708

Mathur, M. B., and Reichling, D. B. (2016). Navigating a social world with robot partners: a quantitative cartography of the Uncanny Valley. *Cognition*. 146: 22–32, 008. doi: 10.1016/j.cognition.2015.09.008

Meadows, D. H. (2015). *Thinking in* Systems: A Primer. Chelsea Green Publishing.

Melo, C. D., Marsella, S., and Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Transactions on Computer – Human Interaction (TOCHI)*. 23:2, 1–17. doi: 10.1145/2890495

Monat, J. P., and Gannon, T. F. (2015). What is Systems Thinking? *A Review of Selected Literature Plus Recommendations. American Journal of Systems Science*. 4:1, 11–26. jss.20150401.02

Moore, C., Detert, J. R., Treviño, L. K., Baker, V. L., and Mayer, D. M. (2012). Why Employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychol*. 65, 1–48. doi: 10.1111/j.1744-6570.2011.01237.x

Mori, M. (2012). [1970]. The uncanny valley. *IEEE Robot. Automat. Magazine*. 19, 98–100. doi: 10.1109/MRA.2012.2192811

Mosier, K. L., Skitka, L. J., Burdick, M. D., and Heers, S. T. (1996). Automation bias, accountability, and verification behaviors. *Proceed. Human Factors Ergon. Soc. Ann. Meet*. 4, 204–208. doi: 10.1177/154193129604000413

Motowidlo, S. J., and Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *J. Appl. Psychol*. 79, 475–480. doi: 10.1037/0021-9010.79.4.475

Müller, J. M., Buliga, O., and Voigt, K. I. (2021). The role of absorptive capacity and innovation strategy in the design of industry 4.0 business Models - A comparison between SMEs and large enterprises. *Eur. Manag. J*. 39, 333-343. doi: 10.1016/j.emj.2020.01.002

Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Edu. Psychol*. 38, 63–71. doi: 10.1207/S15326985EP3801_8

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Factors*. 39, 230–253. doi: 10.1518/001872097778543886

Pereira, V., Hadjielias, E., Christofi, M., and Vrontis, D. (2023). A systematic literature review on the impact of artificial intelligence on workplace outcomes: a multi-process perspective. *Human Resource Manag. Rev*. 33, 100857. doi: 10.1016/j.hrmr.2021.100857

Ramage, M., and Shipp, K. (2009). *Systems Thinkers*. London: Springer United Kingdom.

Ryan, M. (2020). In AI We trust: ethics, artificial intelligence, and reliability. *Sci. Engin. Ethics* 6, 1–19. doi: 10.1007/s11948-020-00228-y

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K. R. (2019). "Towards explainable artificial intelligence," in *Explainable AI, LNAI 11700*, ed. W. Samek et al. (Switzerland: Springer Nature), 5-22. doi: 10.1007/978-3-030-28954-6_1

Senge, P. M. (1990). *The Fifth Discipline*. The Art and Practice of The Learning Organization. Currency. New York.

Sterman, J. D. (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World (51 print)*. Irwin/McGraw-Hill.

Teubner, T., Adam, M., and Riordan, R. (2015). The impact of computerized agents on immediate emotions, overall arousal and bidding behavior in electronic auctions. *J. Assoc. Inform. Sys*. 16, 838–879. doi: 10.17705/1jais.00412

Wagner, A. R., Borenstein, J., and Howard, A. (2018). Overtrust in the robotic age. *Commun ACM*. 61, 22–24. doi: 10.1145/3241365

Wilkens, U. (2020). Artificial intelligence in the workplace—A double-edged sword. *Int. J. Inform. Learn. Technol*. 3, 22 doi: 10.1108/IJILT-02-2020-0022