Technological Forecasting & Social Change 226 (2026) 124491

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Sensemaking and AI: Unraveling individuals' reactions to the black box in a three-study investigation

Domenico di Prisco [a,*], Silvia Dello Russo [b]

[a] IESEG School of Management, 3 rue de la Digue, 59000 Lille, France
[b] Luiss University & Luiss Business School

ARTICLE INFO

Keywords:
Artificial intelligence
Experiment
AI opacity
Sensemaking theory
Human-AI interaction
Augmentation

ABSTRACT

Artificial intelligence (AI) technologies promise to transform how people perform tasks and make decisions within organizations. Yet, their impact on human reasoning processes remains poorly understood. When encountering an unexpected AI suggestion, individuals may either attempt to understand the reasoning behind it or blindly accept or reject it. What drives these different reactions, however, remains unexplored. Unpacking these factors is essential to advance our understanding of augmentation and prevent major decision-making failures. This study addresses this gap through three experimental studies. In study 1 we find that, when performing a task, the unexpected failure of one's own frames increases the likelihood of individuals blindly accepting AI suggestions and effortfully trying to explain them. In study 2 we shed light on the underlying reasons for the results, by analyzing qualitative insights. We find that the unexpected failure of frames promotes "problematization pivoting", a phenomenon wherein individuals anchor their reasoning to opaque AI suggestions ignoring other available cues. In study 3, we add evidence of potential negative performance implications associated with effects documented before. Overall, these findings contribute to the literature on human-AI augmentation and sensemaking theory, while also alerting managers and policymakers on the perils associated with AI use.

1. Introduction

AI technology is increasingly used within organizations to support human cognition. This trend is aligned with the so-called "augmentation perspective" (Keding and Meissner, 2021; Daugherty and Wilson, 2018; Davenport and Kirby, 2016a, 2016b), which supports the idea that human and AI reasoning should be regarded as complementary poles collaborating with each other rather than opposing forces.

Depicting human-AI reasonings as complementary supports the possibility of overcoming human constraints and biases through the formal reasoning and immense processing power of AI technologies (Wang et al., 2020; Balasubramanian et al., 2022), as well as improving the decision models of these systems by exposing their learning algorithm to human tacit knowledge and social awareness (van den Broek et al., 2021). This position is supported by experimental evidence (Fügener et al., 2022; Jussupow et al., 2021; Zhang et al., 2020), which indicates that human-AI collaborations outperform each of these entities alone. Nonetheless, other works also show that it is not easy to achieve an effective integration of the two poles (Lebovitz et al., 2022; Jussupow et al., 2021).

One of the main criticisms linked to the adoption of AI technologies is the opacity of their processes (Burrell, 2016), meaning that the motivations behind their outputs remain obscure to their users. Various explainable AI (XAI) techniques have been developed over the years to enhance the perceived transparency of AI decisions (Haque et al., 2023; Longo et al., 2024; Pumplun et al., 2023). Yet, XAI techniques often have reduced effectiveness when applied to novel AI models, especially in the field of LLM, because of their high complexity (Longo et al., 2024). As a result, organizations may choose not to use more explainable AI models, considering the trade-off between performance and interpretability (Assis et al., 2025) and the fact that XAI models may even have a detrimental effect on human cognition (Bauer et al., 2023). Thus, it is increasingly important to integrate research on XAI methods with behavioral insights into how users respond to algorithmic opacity and what factors drive desired response patterns (Lebovitz et al., 2022). From a user perspective, the issue of opaque AI suggestions is further aggravated by the differences between human and AI rationalities. Such differences are often reported to lead to AI suggestions that are

* Corresponding author.
E-mail address: d.diprisco@ieseg.fr (D. di Prisco).

https://doi.org/10.1016/j.techfore.2025.124491
Received 22 October 2024; Received in revised form 28 September 2025; Accepted 7 December 2025
Available online 6 February 2026

unexpected and for which finding a plausible interpretation is challenging (e.g. suggestions of medical diagnosis beyond the options that a physician was considering) (e.g., Lebovitz et al., 2022; Lu and Zhang, 2025; van den Broek et al., 2021).

In these situations, which are the focus of the present study, existing research shows how human actors may react through either engaged or unengaged forms of augmentation (Lebovitz et al., 2022). Engaged augmentation refers to cases where individuals actively make efforts to comprehend the underlying reasoning behind AI suggestions and, based on that, thoughtfully choose how to incorporate them in their judgments; unengaged augmentation encompasses instances where individuals blindly conform to or refuse AI suggestions. Engaged forms of augmentation include more systematic reasoning to integrate AI judgments in human reasoning processes, thereby requiring more time to be implemented and undermining the efficiency assumptions that underlie many AI applications (Lebovitz et al., 2022). Conversely, unengaged forms of augmentation (Lebovitz et al., 2022) maximize efficiency but expose organizations to risks of extreme failures (in case of unengaged acceptance) or perpetuating existing human biases despite making investments in new technologies (in case of unengaged refusal). For instance, Anthony (2021) shows how unengaged acceptance in response to AI suggestions may lead bankers to give wrong advice to clients for major corporate transactions. Moreover, cases such as Amazon's AI recruitment system, which was found to disadvantage women during the selection process, show how such instances may also perpetuate discrimination (Mujtaba and Mahapatra, 2019). Finally, unengaged refusal may lead individuals to not use, or overrule algorithms (Waardenburg et al., 2022; Cruz, 2024).

Although these insights have shed light on the challenges associated with augmentation, the extant literature lacks studies investigating which factors promote engaged or unengaged forms of augmentation. Collectively, studies offer insights into the individual, technological, and contextual factors that may increase trust in AI (e.g. Renz et al., 2024; Omrani et al., 2022), but they do not examine what factors lead an individual to manifest a specific augmentation type and do not take into account the situated, subjective cognitive conditions under which human-AI interactions take place. Moreover, there is a lack of studies that provide a more detailed understanding of what each augmentation type implies for human reasoning. Expanding knowledge in this area is important not only for theory but also for practice. AI technologies, especially with the recent LLM advancements, are increasingly being used to support organizations in sensitive domains for decision-making such as hiring, strategic decision making, and medicine. Understanding what leads to either engaged or unengaged responses may help prevent the reinforcement of bias, avoid major failures (Lu and Zhang, 2025), and assist organizations in identifying how to prevent or reduce resistance to the use of technology more broadly (Cruz, 2024; Lebovitz et al., 2022).Therefore, coherently with several recent calls in the literature (Dwivedi et al., 2023), the present study aims to answer the following research questions:

Under which conditions do individuals commit to different types of augmentation, and how does exposure to unexpected AI suggestions influence their reasoning?

We examine these questions by highlighting how individuals rely on cognitive frames to overcome their limited information processing capabilities, filter what they notice, construct stable interpretations of the world, and guide their actions (Cornelissen and Werner, 2014). Consequently, suggestions conflicting with established frames are ignored unless individuals meet unexpected failures in these cognitive structures, prompting them to shift their reasoning towards novel cues that would otherwise be overlooked (Maitlis et al., 2013; Weick, 1995). To examine the role of such frames in individuals' responses to AI suggestions, we conducted three experimental studies based on a cognitive task. Two studies offer quantitative tests of the relationships of interest, and one leverages the qualitative analysis of think-aloud protocols (Jussupow et al., 2021; Ericsson and Simon, 1980) to provide a more granular understanding of the mechanisms we postulate.

Our work makes significant contributions to theory and practice. First, we contribute to the literature on human-AI augmentation by showing that individuals are more inclined to consider AI suggestions and blindly conform to them when they experience unexpected failures in their cognitive frames, in line with sensemaking theory (Weick, 1995). Second, through the use of think-aloud protocols, we unpack and provide evidence on each augmentation type, as well as on the factors that account for the results that emerged from the quantitative analysis. In doing so, we contribute to sensemaking theory by introducing the concept of "problematization pivoting". This phenomenon arises when individuals shift their cognitive focus away from the initial task and instead direct their efforts exclusively towards unraveling the reasoning underlying the black-boxed cues, actively searching for plausible explanations that justify the validity of the AI suggestions. We show that problematization pivoting leads individuals to become entirely absorbed in AI recommendations, neglecting the original task. Finally, in the third study, we provide evidence on the potential performance implications of these mechanisms in cases of incorrect AI suggestions. Taken together, these findings alert policymakers and decision-makers to situations in which AI suggestions may reinforce existing biases or result in significant failures.

2. Theoretical background

2.1. Humans and AI: the augmentation perspective

Advances in AI technologies have enabled their adoption to support human actors in a wide range of increasingly complex cognitive tasks, which are those primarily requiring reasoning, problem-solving, or decision-making for their accomplishment. Examples include AI applications for cancer diagnosis (Lebovitz et al., 2022), the prediction of future crimes (Waardenburg et al., 2022), and R&D investment decisions (Keding and Meissner, 2021). The logic underlying human-AI collaborations, lies in the principles of the augmentation perspective (Davenport and Kirby, 2016a, 2016b; Lebovitz et al., 2022; Raisch and Krakowski, 2021), which considers AI and humans as reciprocal enhancers (Baer et al., 2025; Daugherty and Wilson, 2018). This is due to the different yet complementary nature of human and AI reasoning and capabilities. AI operates on patterns synthesized within their decision models and computed through statistical operations on data collected via sensors, external sources, or provided by developers. Unlike humans, AI is not constrained by attention limitations or cognitive capacity and can incorporate a larger amount of data into its reasoning (Krakowski, 2025; Raisch and Fomina, 2025). By contrast, human reasoning is bounded by cognitive constraints, anchored in prior knowledge, and guided by heuristics designed to minimize cognitive effort (Cyert and March, 1963; Krakowski, 2025; Simon, 1957). On the flipside, human reasoning benefits from intuition, tacit knowledge, and context-awareness. Given these differences, it is not surprising that most studies report that AI technologies frequently provide suggestions that diverge from the decision outputs of humans (e.g. Lebovitz et al., 2022; Lu and Zhang, 2025). The potential for augmentation lies in this diversity, so that when the two are successfully integrated the performance of human-AI collaborations on shared tasks surpasses the outcomes achieved by either humans or AI alone (e.g. Fügener et al., 2022; Zhang et al., 2020).

Despite these opportunities, existing research indicates that realizing such synergies may be challenging due to the opacity of AI reasoning processes (Jussupow et al., 2021; Lebovitz et al., 2022; Van Den Broek et al., 2021). Such opacity is especially critical in cases of human-AI disagreement, which are the focus of this study. Indeed, when human and AI judgments align, opacity is not problematic for users, as they refer to AI output merely to confirm their reasoning (Jussupow et al., 2021). In contrast, different user reaction types (i.e. augmentation patterns) may occur in case of human-AI disagreement, namely engaged

and unengaged augmentation patterns (Lebovitz et al., 2022).

Engaged augmentation occurs when individuals actively engage in validation practices to assess the reliability of AI suggestions and understand their underlying rationale. Although these practices can improve decision quality, they also introduce inefficiencies due to the time they require. Unengaged augmentation involves either blind rejection or uncritical acceptance of AI suggestions. On one side, thoughtless refusal of AI suggestions may reinforce existing biases and generate sunk costs. On the other hand, blind acceptance may lead to discrimination, progressive deskilling, and increased risks of extreme failures (Krakowski, 2025; Lebovitz et al., 2022; Raisch and Krakowski, 2021). Empirical research shows that in case of human-AI disagreements, users of AI cope with the opacity of AI suggestions largely through unengaged responses, either by over-relying on or distrusting them without critical evaluation (Lu and Zhang, 2025; Jacovi et al., 2021).

To address the opacity challenge, recent developments on explainable AI (XAI) have introduced techniques aimed at clarifying the relationship between input and output (Haque et al., 2023; Longo et al., 2024; Lu and Zhang, 2025). Examples include feature attribution, which highlights the attributes most critical for a specific model decision, and decision trees, which represent decision rules in a tree-like structure. Explanations provided by XAI methods may adopt different formats (i.e. textual, visual, auditory, or hybrid), focus on clarifying a single output (local explainability) or the functioning of the entire model (global explainability), and be introduced either before or after model training (Haque et al., 2023; Longo et al., 2024; Martens et al., 2025).

Some studies have shown that the adoption of XAI methods is necessary for the integration of AI explanations into human judgment in case of human-AI disagreements (Bauer et al., 2023; Lu and Zhang, 2025), partly because they lead to increased trust towards AI (Schuetz et al., 2025; Glikson and Woolley, 2020). Nevertheless, other studies also indicate that XAI does not fully resolve problems caused by the opacity of AI, due to both technological limitations and the ways in which users respond to explanations. First, authors highlight a trade-off between performance and interpretability (Assis et al., 2025), which discourages the adoption of transparent models. Second, XAI techniques may be less effective with new generative AI models, given their increased complexity, and could therefore increase the risk of derailing users rather than providing support (Longo et al., 2024). From a user perspective, existing research reports that explanations may induce cognitive overload (Haque et al., 2023) and that users tend to draw information only from explanations that align with their preexisting mental models (Bauer et al., 2023).

These insights suggest that, although useful, research on XAI must be complemented by approaches that consider the social aspect of human-AI interactions to realize the potential for their augmentation. In this regard, research streams on trust and technology offer indications on which factors may shape individual reactions to AI suggestions. Elements such as expertise, sector of application, a person's ethical considerations, concerns about discrimination, perceived uncertainty, accountability, and availability of support in case of complaints influence individual trust towards AI (Cruz, 2024; Omrani et al., 2022). At the demographic level, men tend to display higher levels of trust and willingness to use AI than women (Renz et al., 2024; Omrani et al., 2022) and, on average, people over 55 tend to exhibit more trust towards AI (Omrani et al. (2022).

Taken together, the literature has advanced our understanding of the factors that may increase the likelihood of trusting AI suggestions. Yet, there are several questions still open. First, the literature does not differentiate between engaged or unengaged forms of augmentation. Second, most of the examined factors are stable features (of technology, individuals, and organizations), and situated subjective cognitive conditions under which human-AI interactions occur have been neglected. Advancing collective knowledge in this area is essential, especially owing to the increasingly unstructured and uncertain and less stable decision-making domains in which AI decision-support mechanisms are being implemented. In this regard, initial evidence the importance of cognitive frames in shaping users' responses to AI suggestions (e.g. Bauer et al., 2023; Cruz, 2024; Jussupow et al., 2021).

2.2. Cognitive frames, sensemaking, and AI suggestions

Unlike AI technologies, human reasoning struggles when dealing with big volumes of data (Krakowski, 2025; Shrestha et al., 2019). As management scholars have highlighted, "*one thing an intelligent executive does not need is totally accurate perception*" (Starbuck and Milliken, 1988, p. 40) because trying to grasp the complexity of the world slows down decisions and hinders action. Sensemaking theory (Maitlis and Christianson, 2014; Weick, 1995) suggests that humans address this issue by adopting frames, which are "*knowledge structures that help individuals to organize and interpret incoming perceptual information by fitting it into already-available cognitive representations from memory*" (Cornelissen and Werner, 2014, p. 7). These frames are developed through personal experiences, social interactions, and cultural influences and can assume various forms. Of relevance here are the cause maps (Weick, 1979, 1988, 1995), which are cognitive frames stored in the form of "if-then" statements influencing which cues we notice, how they are interpreted, and provide expectations associated with the future outcomes of each action (Cornelissen and Werner, 2014).

According to Weick (1995), organizational actors develop a meaningful understanding of their environment by linking present emerging cues to pre-existing frames. If they are not ignored by perceptual filtering systems (Starbuck and Milliken, 1988), unexpected cues that are incompatible with the existing cognitive structures lead to more time-consuming reasoning patterns, as individuals will try to make sense of them in ways consistent with prior frames (Weick, 1995). Applying these insights to human-AI collaborations to perform a cognitive task, unexpected algorithmic suggestions may extend the time spent on the task as individuals will attempt to reconcile these cues with their existing cognitive structures.

HP1. While performing a cognitive task, subjects exposed (vs. not exposed) to an unexpected AI suggestion will spend more time completing the task.

Frames enable action, but also create blind spots (Cornelissen et al., 2014; Weick, 1988). For instance, policymakers may overlook information that could predict financial crises (Abolafia, 2010). As a result, the pursuit of simplification through frames can also lead to their failure, leaving individuals perplexed and disoriented (Maitlis and Sonenshein, 2010; Weick, 1993). These emotionally charged instances prompt individuals to narrow their attention and engage in systematic and time-consuming reasoning (Weick, 1995). In this process, they will look for cues that may explain the unexpected shock (Maitlis et al., 2013) to reduce their perceived ambiguity.

In this process, we contend that individuals may anchor their sensemaking process to AI suggestions (Maitlis et al., 2013). Several factors support this claim (Logg et al., 2019). First, widespread ideas of AI technologies as hyperrational agents act as signals about their reliability (Appio et al., 2025; Keding and Meissner, 2021). Furthermore, the accuracy scores of these technologies may be perceived as superior to the judgmental confidence of individuals after the experienced shock (Jussupow et al., 2021). Hence, in case AI suggestions are misaligned with the disrupted frames, anchoring sensemaking to these opaque cues will extend the length of individuals' reasoning processes. Without external suggestions, individuals will search for cues that provide explanations close to their frames (Weick, 1995). Conversely, looking for explanations underlying opaque AI suggestions inconsistent with prior frames will necessitate search efforts beyond established mental structures, prolonging the time of the process.

HP2a. While performing a cognitive task, subjects exposed (vs. not

exposed) to a failure in their frame will spend more time completing the task.

HP2b. While performing a cognitive task, the effect of exposure (vs not exposure) to an unexpected AI suggestion on task completion time will be positively moderated by whether subjects experience (vs. not experience) a failure in their frame.

The status of cognitive frames has further implications for how people interact with AI suggestions. In the absence of unexpected failures, individuals will try to resolve the ambiguity of opaque cues by creating interpretations confirming their previous mental structures to avoid the cognitive costs of questioning themselves and their stable interpretations of the world (Maitlis et al., 2013). Under such circumstances, failing to make sense of the AI suggestion through existing cognitive structures will lead individuals to be more inclined to uphold their frames and reject the AI suggestions, labeling them as the result of technological malfunctioning without seeking alternative explanations for them (i.e., *unengaged forms of refusal*).

HP3. While solving a cognitive task, subjects exposed to an unexpected AI suggestion who do not experience (vs. experience) a failure in their frame will be more likely to blindly reject the AI suggestion (i.e. unengaged refusal).

On the other hand, the unexpected failure of frames induces individuals to engage in "sensedemanding" (Maitlis and Christianson, 2014), which entails active efforts to acquire and process information from other human or non-human actors in order to reduce the ambiguity caused by the shock. In such situations, individuals are more susceptible to being influenced by external cues that they would have, otherwise, overlooked. For example, in the case of the Air France 447 flight disaster in 2009 (Oliver et al., 2017), the reasoning of the pilots was derailed by erroneous indications from the flight directors, despite their training to disregard such indications in unreliable speed situations. Building on this, we propose that, when provided with an AI suggestion after an unexpected failure, human actors are less inclined to blindly refuse it (unengaged refusal). This is because the failure will enhance their propensity to anchor their sensemaking process to the opaque AI suggestion to develop plausible narratives that validate or disconfirm it (i.e., *engaged forms of augmentation*). Although a plausible explanation justifying the AI suggestion may not be found, we anticipate that the high performance expected of AI technologies applications (Appio et al., 2025; Keding and Meissner, 2021), and the confusion stemming from the previous failure, will make individuals also more inclined to blindly conform to AI (*unengaged acceptance*). Consequently, the inability to comprehend AI outputs can give rise to algorithmic forms of pluralistic ignorance (Weick, 1990), where individuals are perplexed by what is happening, but assume that no one else is.

HP4. While performing a cognitive task, subjects exposed to an unexpected AI suggestion, who experience (vs. do not experience) a failure in their frames, will be more likely to accept the AI suggestion.

HP5. While performing a cognitive task, subjects exposed to an unexpected AI suggestion who experience (vs. do not experience) a failure in their frames, will be the least likely to respond by displaying an augmentation type of unengaged refusal (vs. other augmentation types).

3. Study 1

To test our hypotheses, we relied on laboratory experiments. This methodology is optimal for addressing research questions involving phenomena that are rare and difficult to observe, such as the effect of AI suggestions after episodes of unexpected cognitive failures (Weick, 1995). Moreover, experiments facilitate investigation by isolating the effects under study (Bolinger et al., 2022), thus minimizing the influence of confounding factors.

3.1. Sample

The sample size was determined using a power analysis by hypothesizing a medium-sized effect (F = 0.25) with an alpha value of 0.05 and a power of 0.8 in a 2 × 2 experimental design, which resulted in a final sample size of 180 participants. The data collection was conducted through Prolific. Considering the higher rates of insufficient effort responding (IER) and attrition associated with the use of online panel data platforms (OPDs) (Aguinis et al., 2021), we aimed for a larger sample size. We collected 312 responses from employed individuals with an education level equal to or higher than bachelor's degree and were native speakers of the researchers' mother tongue.[1] Our screening criteria (Table S1) excluded 25.65 % of the respondents. The duration of the study (mean = 10.54 min; SD = 7.41 min), explains why the IER rate was higher than the one (15–20 %) reported in previous studies (Fleischer et al., 2015). After filtering out such cases, our sample consisted of 232 participants.

Further investigation revealed that 39 participants chose not to view the algorithmic suggestion and were thus excluded from our analysis. A binary logistic regression showed that there were no significant differences in any of the control variables between the participants who chose to view the suggestion and those who did not. The final sample size was 193 participants (50.3 % male). On average, participants were 33 years old and primarily employed in full-time positions (74 %).

3.2. Experimental design and procedure

3.2.1. Experimental task

Two factors were manipulated: exposure to an AI suggestion and the experience of an unexpected failure of previous frames. The group exposed to the AI suggestion was informed that the tool had achieved a 70 % accuracy score in prior tests. The experimental task involved participants solving four logical puzzles based on pattern recognition (Wason, 1960; Evans, 2016). They had to identify the number(s) that complete each pattern and explain the reasoning underlying their solution(s). After submitting the explanation, participants received feedback (positive or negative) on the correctness of their solution. The feedback was manipulated to operationalize the concept of "unexpected failure of prior frames". This task can be considered as cognitive, since its resolution requires capabilities such as reasoning, analysis and problem-solving. Furthermore, we chose this task for three reasons: (1) logical puzzles have narrower decision space compared to other tasks, which facilitated the formation of similar cause maps among participants; (2) logical puzzles represent ambiguous problems, making the manipulation of feedback credible; (3) numeric puzzles encourage using similar reasoning across different rounds, so feedback for one puzzle influences the resolution strategy for the next.

The design of the experimental task (See Fig. 1) was pretested through a pilot study. In the first two rounds, we presented two puzzles with easily identifiable solutions to influence participants' understanding of what constitutes a pattern. This priming fostered the expectation that solving the puzzles required an increasing monotone function (i.e. participants' task-related frame of reference). After solving the first two puzzles, all participants received positive feedback. The third puzzle (in which participants had to identify two numbers) was designed to be solvable using the same reasoning used before, but ambiguous enough to make the feedback manipulation credible. After submitting their answers, half of the participants received positive feedback confirming their frames, and the other half received unexpected negative feedback (a warning on the screen informing them that their solution was correct or incorrect). Finally, in the fourth puzzle, the participants encountered a puzzle that, like the third, could be resolved by applying the same logic adopted to solve the first two puzzles. In this instance, half of the

[1] Anonymized for the review process

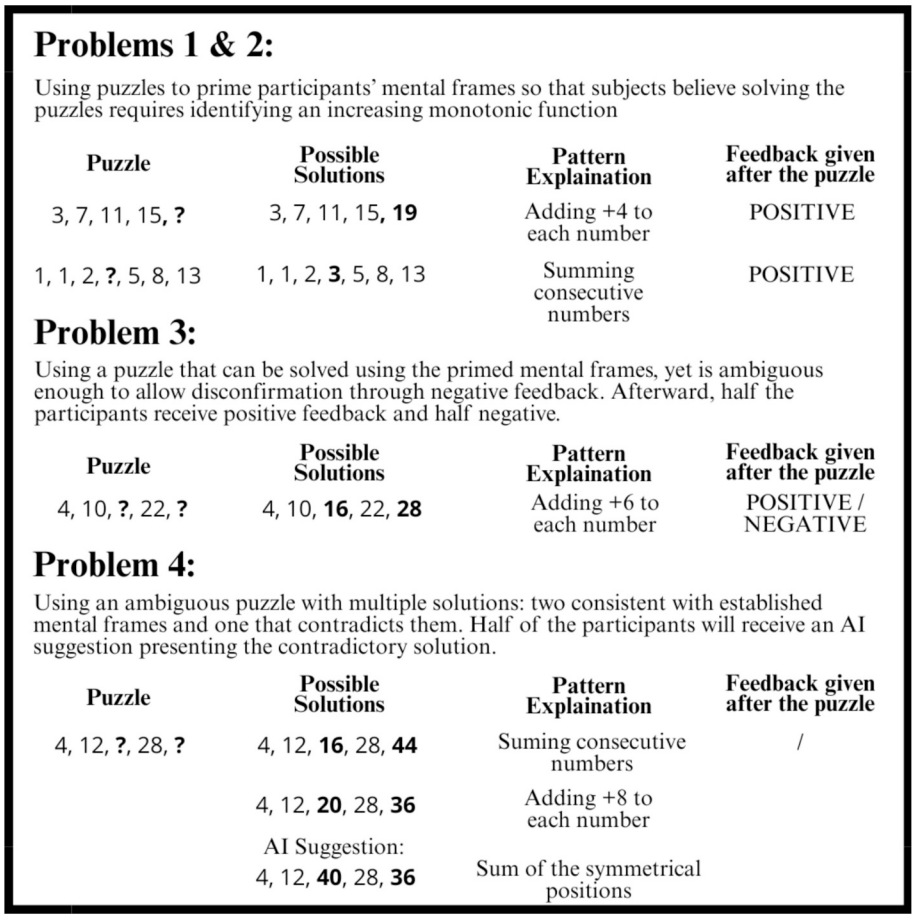

Fig. 1. Summary of the experimental task for Study 1.

participants were exposed to an AI suggestion for resolving the puzzle and half did not. The AI suggestion was deliberately designed to be incompatible with the frames built throughout the previous exercises (that is, not explainable through an increasing monotone function). We chose to make the AI suggestion accessible only through an intentional button click, reflecting the human-AI interaction patterns commonly observed in organizations, where the display of the suggestion is contingent on the practitioners' deliberate intention to view it (Lebovitz et al., 2022). While Puzzles 1–3 were preparatory, the fourth puzzle represented the task used to test our hypotheses.

Upon completing the puzzles, participants compiled a short survey, including attention checks, manipulation checks and control variables.

3.2.2. Rewarding

To discourage IERs we implemented a specific reward system based on multiple components. The first component, equivalent to $2.21 (equivalent to $8.86/h), represented a base reward for correctly completing the study, regardless of the answers provided to the puzzles. The second component, with a maximum value of $1.32 ($5.26/h), was described to be assigned based on participants' ability to provide correct answers accompanied by plausible explanations in the last two puzzles (maximum $0.44 for the third puzzle and maximum $0.89 for the fourth). While the base pay alone was aligned with the "low" level reward standard set by Prolific, achieving the full value of the bonus would elevate it to a "high" level. To mitigate social desirability bias and the careless acceptance of AI suggestions solely to achieve the bonus, we informed participants that the additional reward for correctly responding to the third and fourth puzzles would be halved if they provided correct answers without being able to explain them. Since the puzzles were ambiguous and had multiple plausible answers, we only ensured that participants provided a logically valid explanation to assign the full monetary reward, employing less restrictive criteria than announced (this was revealed in the debrief). On average, each participant received a reward of $4 ($19.24/h).

3.3. Measures

3.3.1. Manipulated variables

AI suggestion is dichotomic and takes a value of 1 if the participants were exposed to the AI suggestion while solving the fourth puzzle. Unexpected failure of frames took the value of 1 if the participants received negative feedback after completing the third puzzle and 0 if they received positive feedback instead.

3.3.2. Dependent variables

We collected three dependent variables (DVs): completion time, suggestion acceptance and augmentation. Completion time, collected for all conditions, is a continuous variable measured as the duration between the start of the fourth puzzle and the submission of a solution for it. Suggestion acceptance is a dichotomous variable that takes a value of 1 if the participants' answer coincides with the external aid provided during the exercise. This was collected only in the conditions that were exposed to the AI suggestion, likewise augmentation type, which is a nominal variable that can take four different values. We operationalized engaged (unengaged) acceptance as cases where participants adhered to the AI suggestion and were able (unable) to explain the logic underlying their answers. Engaged and unengaged refusal referred to responses that diverged from the AI suggestion; we distinguished between these two augmentation types by asking participants to select one of three statements that best described their reasoning. Unengaged refusal cases

selected one of the following statements: *"I ignored the suggestion after a few seconds because it immediately felt wrong to me"* or *"I ignored the suggestion from the start because I didn't want any help"*. In contrast, engaged refusal cases chose the statement: *"Although I invested a considerable amount of time trying to comprehend the suggestion, I ultimately decided to ignore it because I could not identify a pattern to justify it."*

3.3.3. Control variables

Given that individuals' reactions to AI technologies may depend on individual characteristics such as personality traits (Mahmud et al., 2022), we controlled for personality traits using a translated version of the Big Five short scale (Gosling et al., 2003). Since conformity to AI may also depend on individuals' general attitude towards technological solutions (Glikson and Woolley, 2020), we also controlled for the propensity to trust technology. This variable was operationalized using three items from the trust in technology measurement scale (Mcknight et al., 2011): *"My typical approach is to trust new technologies until they prove to me that I shouldn't trust them," "I usually trust a technology until it gives me a reason not to trust it,"* and *"I generally give a technology the benefit of the doubt when I first use it."* Furthermore, participants self-assessed their skill in solving logic puzzles before starting the first exercise by indicating their level of agreement (1–5) with the statement: *"I have a good knowledge of both Logic Puzzles and the methodologies to solve them (e.g., find the missing number of the numeric series)."* Lastly, we collected demographic information from the participants, including their gender, age, education, and employment type (full-time/part-time).

3.4. Results

The descriptive statistics and intercorrelations among variables are reported in Table 1.

On average, the completion time of the fourth puzzle was 88.41 s. The subjects exposed to both negative feedback and AI suggestion showed the highest mean completion time (124.27 s) while the control group without any of the two manipulations was the fastest at providing an answer (46.24 s). Among the subjects exposed to the AI suggestion (n = 93), the prevalent augmentation type was engaged refusal (n = 35, 37.6 %) while cases of engaged acceptance were the least frequent (n = 3, 3.2 %). No one in the control group autonomously identified the solution we used for the AI suggestion.

Very few correlations among our DVs and control variables were significant. Completion time was positively associated with openness, while the augmentation type of engaged refusal was positively associated with the level of education. For this reason, and to improve the interpretability and external validity of our findings (Becker et al., 2016), we decided to present the results of our analyses without control variables (The results including control variables are available from the first author upon request).

To test Hypotheses 1, 2a and 2b, we conducted an analysis of variance (ANOVA). The findings support HP1, highlighting a significant difference ($F(1,190) = 12{,}809$, $p < .001$, $\eta^2 = 0.063$) in the completion time of those who were exposed to the AI suggestion (M = 109.68 s) versus those who were not (M = 68.62 s). We also found support for HP2a by detecting a significant difference ($F(1,190) = 13{,}075$, $p < .001$, $\eta^2 = 0.064$) in the completion time of those who experienced an unexpected failure of their prior frames (i.e., received negative feedback) versus those who did not experience such failure (i.e., received positive feedback; M = 107.96 vs. M = 66.49). Finally, the interaction between the exposure to AI suggestions and the experience of prior frame failures on task completion turned out to be non-significant ($F(1,189) = 0.288$, p = .592), and therefore the hypothesis 2b is not supported.

Hypotheses 3, 4, and 5 were tested on the subsample (n = 93) of subjects who were exposed to the AI suggestion (i.e., two experimental conditions). To test Hypothesis 3, we ran a binary logistic regression. The logit model correctly classified 69.9 % of the cases. Its fit with the

Table 1
Descriptive statistics and correlations among Study 1 variables.

		M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	Completion time	88.41	79.84																			
2	Suggestion acceptance (0/1)	0.29	0.47	0.24*																		
3	Unengaged refusal (0/1)	0.33	0.47	−0.21*	−0.45***																	
4	Unengaged acceptance (0/1)	0.26	0.44	0.2	0.92***	–																
5	Engaged refusal (0/1)	0.38	0.49	−0.02	−0.50***	–	–															
6	Engaged acceptance (0/1)	0.03	0.18	0.13	0.29**	–	–	–														
7	AI suggestion (0/1)	0.48	0.5	0.26***	0.00	–	–	–	–													
8	Unexpected failure of prior frames (0/1)	0.53	0.5	0.26***	0.23*	−0.38***	0.23*	0.15	0.04	0.06												
9	BFExtraversion	4.03	1.19	−0.00	0.09	−0.04	0.1	−0.05	−0.01	0.01	0.03											
10	BFAgreeableness	5.07	1.07	0.07	0.05	−0.04	0.02	−0.00	0.08	−0.02	−0.21**	0.00										
11	BFConscientiousness	5.21	1.27	0.02	0.01	0.12	−0.02	−0.13	0.07	−0.01	−0.07	0.60***	0.09									
12	BFEmotionalstab	4.36	1.5	−0.02	−0.12	0.15	−0.19	−0.03	0.18	−0.1	−0.18*	0.29***	0.22**	0.39***								
13	BFOpenness	4.61	1.13	0.19**	0.19	−0.02	0.16	−0.16	0.09	−0.06	−0.02	0.11	0.17*	−0.05	0.09							
14	Propensity to trust Technology	3.6	0.73	−0.06	0.11	−0.14	0.04	0.04	0.18	0.07	0.01	0.12	0.02	−0.04	0.04	0.04						
15	Age	32.59	8.57	0.08	0.04	0.14	0.03	−0.17	0.03	0.02	−0.08	0.1	0.12	0.21**	0.23**	0.07	0.06					
16	Gender (male = 1, female = 2)	1.50	0.50	0.06	0.08	−0,11	0.10	0.03	−0.06	−0.01	0.01	0.07	0.02	0.01	−0.19**	0.08	−0.04	−0.14				
17	Education	1.82	0.6	0.03	−0.12	−0.13	−0.15	0.24*	0.06	0.02	−0.04	0.09	0.02	0.13	0.04	−0.03	−0.13	0.19**	−0.12			
18	Employment type	1.74	0.44	0.02	−0.04	−0.03	−0.1	0.07	0.12	−0.09	0.01	−0.01	−0.07	0.03	0.19**	−0.11	−0.02	0.13	−0.29***	0.20**		
19	Task skill	3.25	0.85	0.05	−0.02	−0.05	−0.04	0.07	0.07	0.11	0.00	0.12	0.11	0.13	0.04	−0.04	−0.06	−0.02	−0.07	0.06	−0.03	

Note. Correlations involving Suggestion acceptance, unengaged refusal, unengaged acceptance, engaged refusal and engaged acceptance are calculated on $N = 93$. The remaining correlations are calculated on $N = 193$. Correlations between unengaged refusal, unengaged acceptance, engaged refusal, and engaged acceptance are not displayed because these variables represent mutually exclusive responses from a single-choice question.
* $p \leq .05$.
** $p \leq .01$.
*** $p \leq 0.001$.

data significantly improved the null model ($\chi2$ (1) = 13.857, p < .001, Nagelkerke R^2 = 0.192). In our sample, receiving positive vs. negative significantly increased the odds of opting for unengaged refusal (β = 1.711, OR = 5.532, 95 % CI [2.150, 14.233], Wald = 12.588, p < .001).

To test Hypothesis 4, we ran a binary logistic regression. The logit model correctly classified 71 % of the cases. Its fit with the data significantly improved the null model ($\chi2$ (1) = 5.283, p < .022, Nagelkerke R^2 = 0.079). The analysis showed a positive relationship between the exposure to an unexpected failure of prior cognitive structures and the odds to conform to AI suggestion (β = 1.110, OR = 3.036, 95 % CI [1.132, 8.144], Wald = 4.86, p = .027), which supports H4.

We tested Hypothesis 5 through a multinomial logistic regression. The model correctly classified 48.4 % of the cases. Its fit was significantly better than the null model (χ^2 (3) = 14,455, p = .002, Nagelkerke R^2 = 0.158). In our sample, the subjects receiving negative vs. positive feedback were more likely to opt for unengaged acceptance (β = 1.992, OR = 7.333, 95 % CI [2.195, 24.501], Wald = 10.48, p = .001) and engaged refusal (β = 1.544, OR = 4.685, 95 % CI [1.650, 13.300], Wald = 8.417, p = .004) than unengaged refusal. No significant difference was observed with respect to the response of engaged acceptance (β = 1.587, OR = 4.889, 95 % CI [0.392, 60.922], Wald = 1.520, p = .218). These findings partially support HP5.

4. Study 2

Study 1 showed that providing individuals with AI suggestions incompatible with their prior frames leads to spending more time to complete the task. It also revealed that negative feedback did not interact with the AI suggestion to affect task duration. This may be due to the mixed effects of unexpected failures, which can trigger both faster reasoning processes, through unengaged acceptance, and slower completion time through engaged refusal. To gain a more fine-grained understanding of these insights, we conducted Study 2. We integrated the previous design with the methodology of think-aloud protocols (Ericsson and Simon, 1980) explicitly aimed at unpacking the reasoning of participants while solving the task.

4.1. Sample

Prior studies utilizing think-aloud protocols have typically employed sample sizes smaller than 20 subjects due to the extensive time required for their analysis (Laureiro-Martinez et al., 2023). We exceeded this numerosity, collecting answers from 69 management master's students. The use of a student sample is appropriate when investigating questions related to cognition (Bolinger et al., 2022). The study was conducted in two different languages. Subjects unable to verbalize sufficiently in either language were excluded from the sample. After excluding these students and filtering out other cases (Table S2), the final dataset consisted of 50 subjects, most participants in the final sample were female (68 %), with an average age of 23 years.

4.2. Experimental design and procedure

The experiment was conducted online with the virtual presence of the principal investigator (rather than asynchronously as in Study 1). Participants were instructed to connect from a silent place, away from potential distractions. The principal investigator was visible only before the start of the study and informed participants that they would have to think aloud during the resolution of the four puzzles. Before the start of the study, a demonstration of the functioning of the tool providing the suggestion was given to remove any doubts about its existence.

4.2.1. Experimental task

The experimental task was the same as in Study 1, with two variations. First, we adopted a one-factor design where we only manipulated the experience of an unexpected failure of previous frames. Second, the display of the AI suggestion was not optional and was programmed to be shown automatically at the start of Puzzle 4.

4.2.2. Think-aloud protocols

The participants vocalized their thoughts during the experiment (Ericsson and Simon, 1980), and their sessions were video and audio-recorded throughout. Before the start of the experiment, participants solved warm-up tasks (multiplications) while thinking aloud, until they reached a sufficient level of verbalization. To reduce potential sources of anxiety, we reassured subjects that there were no time constraints. We also conducted the experiments remotely by asking participants to share their screens to allow them to remain in a familiar environment.

Upon completing the puzzles, participants were asked to fill in the same survey included in Study 1, with the addition of one more manipulation check asking participants if they believed the external suggestions came from an AI technology.[2] Before the final debriefing, the principal investigator conducted semi-structured interviews, which explored the strategies used to solve the puzzles, the reasoning processes and perceptions of participants during the fourth exercise. Virtual observation and interviews were crucial to complement think-aloud protocols (Jussupow et al., 2021), addressing challenges in verbalizing thoughts during intense cognitive processing (Ericsson, 2003).

4.2.3. Rewarding

In this study, individual rewards were replaced by a lottery system offering monetary vouchers as prizes (with a total value equal to US $208). Participants were informed that their inability to provide plausible explanations for the answers submitted for Puzzles 3 and 4 decreased their chances of victory.

4.3. Measures

4.3.1. Independent variables

The unexpected failure of prior frames has been manipulated as negative feedback as in Study 1.

4.3.2. Dependent variables

The acceptance of AI suggestion and completion time were operationalized as in Study 1. Based on the qualitative analysis of think-aloud protocol, the augmentation type of each participant was identified as follows. (Un)Engaged acceptance patterns indicated instances where subjects (did not) identified a plausible explanation for the AI suggestion before accepting it. Engaged (unengaged) refusal patterns involved cases where the subjects effortfully (did not or effortlessly) tried to explain the AI suggestion before rejecting them. Following these categories, the first author and a research assistant independently coded the augmentation pattern displayed by each participant. The comparison of the final codes determined a simple intercoder agreement of 96 % and a Cohen's kappa of 0.94 (Cohen, 1960), indicating a high level of agreement between coders.

Study 2 also includes an additional DV, namely the time dedicated to each phase within the problem-solving process. This was derived by coding think-aloud protocols using the problem-solving model developed by Laureiro-Martinez et al. (2023).

4.3.3. Control variables

The same as in Study 1, except for education level and employment type.

[2] It was not utilized in Study 1 because, in absence of interviews and observations, we could not distinguish cases where participants rationalized the AI suggestion as a manipulation attempt due to their inability to identify a plausible pattern explaining the algorithmic cues.

4.4. Quantitative and qualitative data analysis

For the quantitative analyses, we ran similar analyses to Study 1 (ANOVAs, binary logistic regression, and multinomial logistic regression).

The recordings of the think-aloud protocols and the follow-up interviews were transcribed separately. Consistent with previous research (Jussupow et al., 2021), our data analysis relied on an abductive approach. To unpack the problem-solving phases, we systematically categorized reasoning processes following the comprehensive model developed by Laureiro-Martinez et al. (2023). The adoption of this framework to analyze the think-aloud protocols is justified by the fact that the puzzles participants had to solve can be conceptualized as problems. In line with prior literature, they represent novel and ambiguous situations that involve a significant gap between the current state and the achievement of relevant goals (i.e., performing well to maximize monetary rewards), thus providing occasions for sensemaking (Weick, 1995). However, we expanded upon this model to better suit the objectives of our study (Table 2 summarizes the framework used for this study). First, we differentiated between frame stating, direction setting, and search phases, and their algorithmic counterparts. This distinction was crucial to identify phases in problem solving that are independent or, rather, anchored to the AI suggestion, involving reading, problematizing, and explaining it, respectively. Second, considering the nature of our task, we replaced the implementation phase of the original model with the search phase, and the implementation evaluation with the search evaluation. Using this framework, the principal investigator coded each segment of the think-aloud protocols and transcribed the timestamps related to each segment from the original recordings. To ensure the reliability of our analysis, a research assistant independently performed the same coding exercise. The comparison of the final codes yielded a simple intercoder agreement of 95.3 % and a Cohen's kappa of 0.94 (Cohen, 1960), indicating a high level of agreement between coders.

By aggregating the temporal duration for each portion of transcription with the same code, we obtained data on the time that each participant spent in each of the problem-solving phases.

Finally, following the prescription of grounded theory (Strauss and Corbin, 1998), we coded and compared the structure and characteristics of the reasoning patterns associated with each augmentation type, obtaining insights into analogies and differences. This iterative analysis was integrated with data extracted from interviews and virtual observations.

Table 2
Problem-solving phase coding definitions and examples.

Phase	Description	Examples of verbalized thoughts (transcribed verbatim)
Frame stating	Verbalizing data mentioned in the text of the problem for the first time (excluding AI suggestion)	"Identify the missing numbers to complete the pattern"
Algorithmic frame stating	Verbalizing data related to the AI suggestion	"so… the AI suggests 12, 40, 28, 36…"
Frame assuming	Enunciation of assumptions related to the problem, usually connected with subjects' mental models	"In theory, these things are usually in ascending order, I believe, and..."
Direction setting	Defining a general search path independent from the AI suggestion	"So I will not look at the suggestion below, I want first to do it by myself and then we'll see"
Algorithmic direction setting	Defining a general search path anchored to the AI suggestion	"I need to understand this 40, I need to understand this relationship between 40 and 4"
Evaluation	Judging search efforts and solutions through evaluations anchored to element not included in the puzzle (e.g. beliefs)	"This thing that you said the AI is 70 % of the times right is like to say… the 30 % of the time it is not, but I don't want to doubt the AI [solution] against mine…"
Decision	Enunciation of final solutions or other decisions related to the problem	"Ok I will go with 20 and 36."
Search	Looking for solution to the pattern independent from the AI suggestion	"4 12 is multiple by 3. 12 to something and then 28. 12 to 28 is… ehm… not multiplication, it's an addition"
Algorithmic search	Looking for plausible explanation justifying the AI suggestion	"let me think a little bit more 4 + 8 is 12, 12 to 40 they added 28. and then… Oh! he went to 28 mmh. And then he went 36."
Search evaluation	Evaluations related to the search efforts or to the solutions based on the data of the puzzle	"So no, but in my opinion, it's not 40, I mean, I don't understand the meaning of 40"

4.5. Quantitative results

The descriptive statistics and intercorrelations among variables are reported in Table 3. The small sample size of this study, justified by the purpose of qualitatively analyzing the behavior of participants, limits the generalizability of the present quantitative findings. Nevertheless, we report the results from the quantitative analysis that are new compared to study 1 (the full replication analysis is available from the first author upon request). As for study 1, we present the results without control.

The descriptive statistics indicate that subjects who received negative feedback ($n = 7$) were more likely to accept the AI suggestion than those who received positive feedback ($n = 4$).

Additionally, the analysis shows that the problem-solving processes of the two experimental groups were statistically different only for the time spent on the algorithmic search phase. Subjects receiving negative feedback spent more time trying to explain the AI suggestion ($F(1,48) = 10.684$, $p = .002$, $\eta 2 = 0.182$) than those who received positive feedback (M = 106 s vs. M = 41 s). This finding supports our hypothesis that the additional time spent by subjects experiencing an unexpected failure of their frames is mostly used to look for plausible narratives justifying the AI suggestions. Fig. 2 provides an overview of the differences between the problem-solving processes of the participants who accepted (or rejected) the AI suggestion. The graphical representation shows that the main difference between the two groups lies in the direction of their search efforts, which are oriented towards the AI suggestion in cases of acceptance and towards independent solutions in cases of rejection.

Fig. 3 offers a more detailed graphical overview of the time spent on each problem-solving phase for each augmentation pattern. An important observation is that, even within the groups of participants who accepted or rejected the AI, there are notable differences depending on the augmentation pattern they displayed. For example, participants following unengaged augmentation patterns devoted little time to the phase of algorithmic search, whereas participants committed to engaged augmentation patterns spent considerable time on it. To clarify the reasons behind these differences, we turn to the qualitative evidence we collected.

4.6. Qualitative findings

Fig. 4 presents the augmentation types enacted in response to the AI suggestions. All the participants responded to the AI suggestion by expressing that "*there was something that did not make sense right from the start*" (Participant 51), revealing a mismatch between prior frames and the AI suggestion. As a result, most participants reported having experienced a feeling of puzzlement. Two main choices distinguish the augmentation patterns they displayed: whether to search for a plausible explanation for the AI suggestion, and whether they ultimately accept the AI suggestion. Moreover, the differences within each augmentation type allowed the identification of sub-patterns, namely strong and weak

Table 3
Descriptive statistics and correlations among Study 2 variables.

		M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	Augmentation type	1.31	1.05																			
2	Completion time	97.18	86.00	0.19*																		
3	Suggestion acceptance (refusal = 0, acceptance = 1)	0.25	0.44	0.31***	−0.09																	
4	Correctness (correct = 1, wrong = 2)	0.86	0.35	0.15	0.08	−0.25*																
5	Suggestion type	2.52	1.11	−0.22*	0.19*	0.16*	−0.21**															
6	Task skill	3.25	0.92	−0.05	−0.11	0.01	−0.00	0.04														
7	Risk aversion	3.11	0.37	0.00	0.03	−0.03	−0.05	−0.01	−0.01													
8	Propensity to trust technology	3.39	0.68	−0.04	−0.12	0.20**	−0.10	0.06	0.14	0.02												
9	Cognitive reflection	2.25	0.88	0.06	0.02	0.09	0.18*	−0.01	0.16*	0.06	−0.13											
10	BF Conscientiousness	5.13	1.11	−0.01	−0.08	0.10	0.05	0.10	0.04	0.20*	0.07	−0.03										
11	BF Emotional stability	4.29	1.33	0.10	−0.06	0.10	0.02	0.02	0.15	−0.05	0.15	−0.02	–									
12	BF Openness	4.70	1.19	0.05	0.03	0.03	−0.10	0.01	0.06	−0.12	0.06	−0.01	–	–								
13	BF Extraversion	3.20	1.51	−0.11	−0.07	−0.01	−0.07	0.00	0.19*	−0.16*	−0.03	−0.05	–	–	–							
14	BF Agreeableness	5.14	1.16	0.00	0.10	0.03	0.01	0.12	0.09	0.10	0.09	−0.01	–	–	–	–						
15	Education	1.77	0.58	0.06	−0.01	0.04	0.02	0.05	−0.09	−0.20*	−0.05	−0.18*	0.01	0.08	0.07	0.17*	0.04					
16	Age	32.84	9.69	0.17	0.08	0.14	0.00	0.06	−0.05	−0.11	0.03	−0.27***	0.08	0.17*	0.04	0.05	0.03	0.23**				
17	Gender (male = 1, female = 2)	1.46	0.50	0.02	0.06	−0.08	0.02	0.03	−0.15	0.12	−0.12	−0.07	0.09	−0.26***	0.32***	0.08	0.09	0.12	−0.12			
18	Employment type	2.35	0.82	0.04	0.02	−0.06	−0.04	−0.00	−0.06	0.00	−0.05	−0.15	−0.01	0.05	−0.02	0.21**	−0.16*	0.22**	0.06	−0.02		
19	Trust in AI	4.08	1.10	−0.02	−0.07	0.15*	−0.16*	0.01	0.11	0.04	0.46**	0.02	0.07	0.15	−0.06	0.04	0.15	−0.05	−0.05	−0.17*	0.12	

Note. Correlations between unengaged refusal, unengaged acceptance, engaged refusal, and engaged acceptance are not displayed because these variables represent mutually exclusive responses from a single-choice question.
* p ≤ .05.
** p ≤ .01.
*** p ≤ 0.001.

versions of each of them. In engaged augmentation types, strong patterns refer to cases in which participants identified both an autonomous solution and a plausible explanation for the AI suggestion, while in weak patterns they identified only one of the two. In unengaged patterns, strong types refer to cases in which no or only superficial attempts were made to explain the AI suggestion, unlike in weak types.

The analysis of both the think-aloud protocols and the interview transcripts led to the identification of several factors explaining why participants displayed different augmentation patterns, namely general attitude towards AI, fear of AI misguidance, self-efficacy towards task-related skills, personality traits, perceived reliability of prior task-related frames, propensity towards self-reliance, and the identification of an explainable solution. Descriptions and illustrative quotes for each of these factors are displayed in Table S3. These factors explain the different behaviors displayed by our participants. More specifically, each factor influenced their levels of self-confidence and the perceived reliability of AI, shaping the decision of 1) whether to search for a plausible explanation of the AI suggestion or not, and 2) whether to accept the AI suggestion or not. Despite the unexpected negative feedback before the fourth puzzle decreased the perceived reliability of prior frames across all participants', we still observe different reactions as other individual differences remain. For instance, individuals displaying a high degree of panic and anxiety after the unexpected feedback, and therefore showing lower emotional stability, were more prone to react to the shock by focusing and ultimately accepting the AI suggestion. In other cases, attempts to explain the AI suggestion resulted in the serendipitous identification of additional, independent patterns, which were eventually accepted.

Despite these individual differences, our analysis also shows that, in most of the acceptance cases (7 out of 11), participants became entirely absorbed in the task of explaining the AI suggestion throughout the exercise until the submission of their final decision, without considering the possible existence of independent solutions. This occurred even when no explanation was found for the AI suggestion. We refer to this as "problematization pivoting". In our analysis, participants were labelled as experiencing problematization pivoting if, once they visualized the AI suggestion, they spent the remaining completion time attempting to make sense of it. This phenomenon helps explain our findings, as it clarifies both the process that led participants to spend increased time on the exercise, particularly on algorithmic search, and their eventual decision to accept the suggestion. As one participant described: "*At the moment when I received external advice that suggested a certain logic behind that exercise, I actually had much more difficulty [...] It was as if I had this distorted vision where I couldn't think... I mean, I had to primarily focus on justifying the artificial intelligence's answer rather than finding a solution on my own*" (Participant 53).

The analysis continues with an outline of the features characterizing each augmentation pattern.

4.6.1. Unengaged refusal

This cluster comprises all participants ($n = 12$) who responded to the AI suggestion by dedicating a few seconds or (in most cases) no time to searching for explanations behind the algorithmic cue. This pattern was predominantly observed in participants assigned to the positive feedback group ($n = 11$ out of the 12 units of the cluster). One participant summarized this process: "*As soon as I saw the 40, I didn't really bother to try and figure out what the artificial machine's reasoning could have been because, for me, the 40 was an unusable answer, it didn't make sense.*" (Participant 44). Based on the different behaviors of these participants, we distinguish between strong ($n = 10$) and weak ($n = 2$) forms of unengaged refusal. In strong cases of unengaged refusal, participants rejected the AI suggestion immediately after reading it, without attempting to provide any explanation, due to its incoherence with prior cognitive frames. In weak cases, participants spent a few seconds trying to make sense of the AI suggestion before rejecting it for the same reasons.

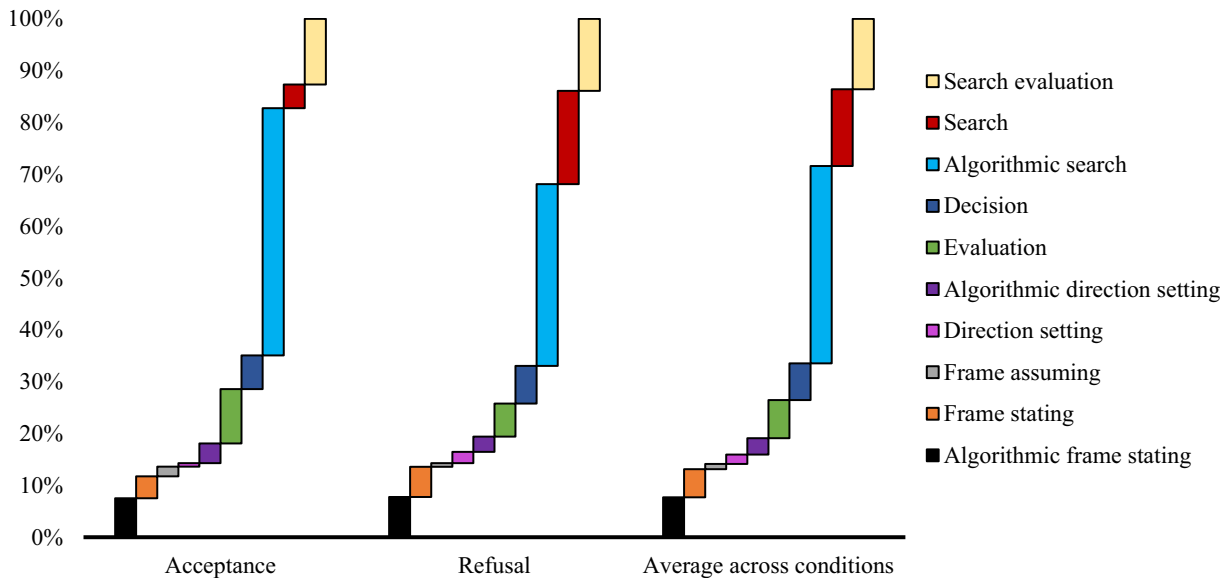

Fig. 2. Average duration (percentage) of the problem-solving phases for average acceptance vs. refusal patterns in Study 2.

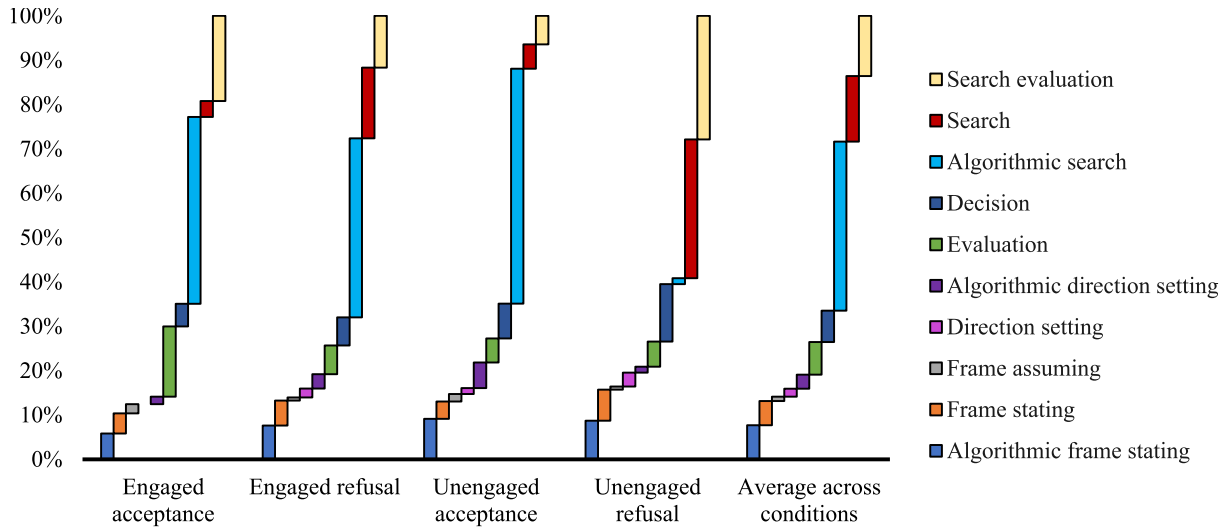

Fig. 3. Average duration (percentage) of the problem-solving phases for each augmentation pattern in Study 2.

4.6.2. *Unengaged acceptance*

This group is the smallest among the augmentation types ($n = 6$), with a higher prevalence observed in the group receiving negative feedback ($n = 5$ out of the six units of the cluster). The key characteristic of this cluster is that all participants ultimately chose to accept the AI suggestion without being able to provide a logical explanation for it, resulting in low levels of confidence in their final answers. Besides this commonality, we identified strong (n = 1) and weak (n = 5) forms of unengaged acceptance.

In case of strong unengaged acceptance, which exclusively encompasses a subject assigned to the negative feedback group, the participant immediately accepted the suggestion, spending no time attempting to explain it: "*Sincerely, I didn't look at the exercise; it [the suggestion] was enough for me to know that it increased the probabilities by 70 percent … I assumed that this exercise would be more difficult than the other one I got wrong, so I trusted the AI*" (Participant 31).

Conversely, participants engaging in weaker forms of unengaged acceptance expressed doubts about the AI suggestion and therefore searched for a plausible explanation for it. Soon, all participants experienced various degrees of problematization pivoting, spending all the remaining time unsuccessfully trying to make sense of the AI suggestion before blindly accepting it. As reported by one of the participants: "*Because I was thinking, 'Okay, maybe the AI is wrong… but it isn't wrong, so I won't try to solve it by myself.' At that moment, I became obsessed with finding the pattern by looking at the solution*" (Participant 51).

4.6.3. *Engaged refusal*

Most participants ($n = 27$) exhibited engaged refusal as their reasoning pattern, with the majority being part of the negative feedback group ($n = 17$ out of the 26 units of the cluster). Despite actively attempting to explain the AI suggestion, participants in this category rejected it. We distinguish between strong ($n = 4$) and weak ($n = 23$) forms of engaged refusal based on success, or lack thereof, in comprehending the reasoning underlying the AI suggestion. In the weak forms of engaged refusal, participants were not able to explain the AI suggestion, ultimately labeling the cue as a malfunction. In the cases of strong forms of engaged refusal, participants were able to independently identify a solution, while also providing plausible explanations for the algorithmic suggestion. As a result, these participants felt puzzled and confused. This ambiguity was resolved by choosing the solution

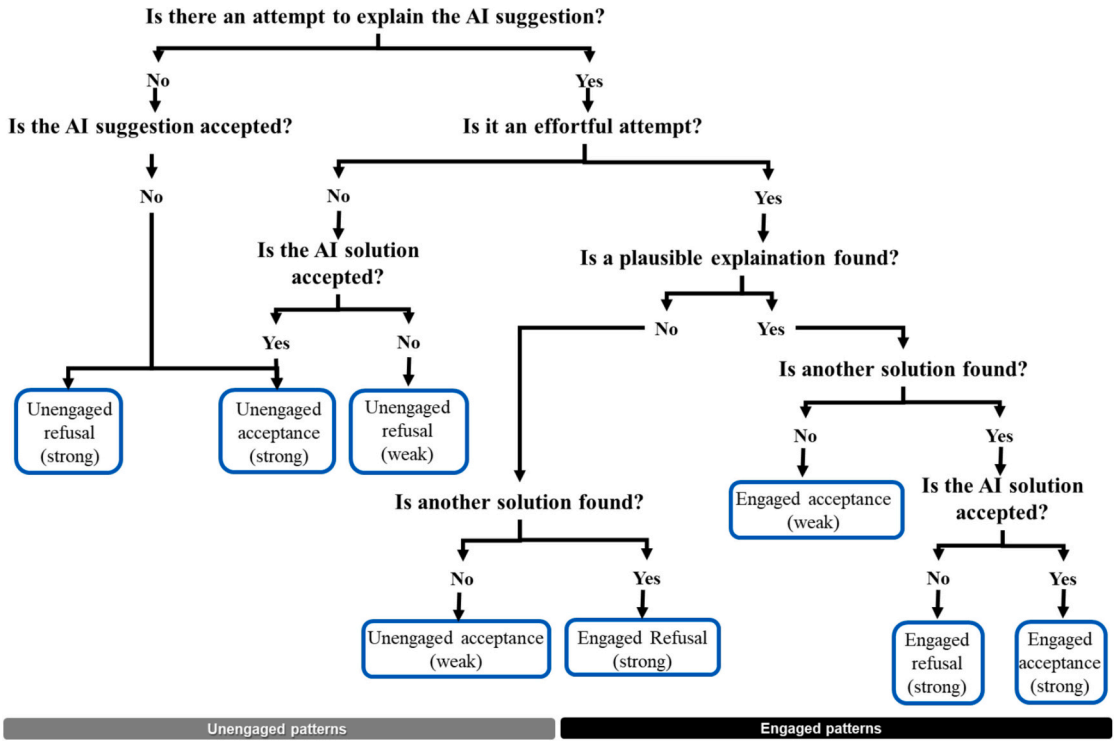

Fig. 4. Synthesis of the augmentation types enacted in response to the AI suggestions in Study 2.

coherent with their prior frames, deemed simpler and more consistent with previous puzzles.

4.6.4. *Engaged acceptance*

A small portion of the subjects in our sample (n = 5), distributed nearly evenly between the negative (n = 2 out of the 5 units of the cluster) and positive feedback (*n* = 3 out of the 5 units of the cluster) groups, accepted the AI suggestion by providing a plausible explanation for it. All the subjects reacted to the AI suggestion by looking for a plausible explanation for it. In contrast to the unengaged group, these subjects succeeded in the task by identifying a pattern that fell outside of their prior frames and accounted for the AI suggestion.

Also in this case, we differentiate between weak (n = 2) and strong forms (n = 3) of engaged acceptance. In weak forms of engaged acceptance, individuals responded to the AI suggestion through problematization pivoting, thus devoting all the remaining time in the exercise to making sense of the opaque algorithmic cue. Differently, individuals who assumed strong forms of engaged acceptance found not only an explanation for the AI suggestion but also an alternative solution to the puzzle, ultimately choosing to accept the former. Upon discovering multiple plausible solutions for the same exercise, these participants felt puzzled and confused. The decision to rely on the AI suggestion depended on the comparison between the confidence in their solution and the AI's accuracy score.

5. Study 3

In the presence of unexpected failures of prior frames, Study 1 showed that participants were more likely to consider and accept AI suggestions that were incoherent with their pre-established frames. It also showed that the incoherent suggestions led to spending more time on the task. Study 2 further explored these results, mapping each augmentation type, conceptualizing the phenomenon of problematization pivoting that partially accounts for these results and highlighting other factors shaping them. However, while conformity, augmentation type, and time are relevant dependent variables, one question remains unanswered, namely to what extent exposure to unexpected AI suggestions, after experiencing a disruption of prior frames, may affect the likelihood of providing correct answers. This gap is particularly relevant given that AI technologies are fallible, prone to hallucinations, and that humans and AI often exhibit poor performance on similar cases (Fügener et al., 2022).

In addition, the type of suggestion provided by the AI may also play a role in users' reactions to AI. Given the condition of failure of cognitive frames, we expect that participants will require significantly less time to construct a plausible explanation for AI suggestions that are coherent with prior frames of reference (in this case, with the frame of monotone increasing patterns) compared to AI suggestions that are incoherent with prior frames. The greater explainability of coherent suggestions is also expected to increase episodes of engaged augmentation and acceptance, as compared to incoherent suggestions.

Considering this, the aim of Study 3 is therefore to investigate the implications of two new variables in our design, namely suggestion type and answer correctness.

5.1. *Sample*

The data collection was conducted through Prolific following the same modalities as in study 1. We collected 203 responses from employed individuals with an education level equal to or higher than a bachelor's degree and sharing the same primary language as the researchers. Our screening criteria (Table S4) excluded 19.4 % of the respondents. After removing these cases, the final sample consisted of 163 participants.

5.2. *Experimental design and procedure*

The experimental task was the same as in Study 1 and it unfolded across 4 puzzles (as summarized in Fig. 1), with a few alterations. In this study, all participants received unexpected negative feedback after puzzle 3, and the independent variable was manipulated across four levels based on the type of suggestion provided. Therefore, study 3 is

based on a 1 × 4 design with four conditions: no suggestion (after puzzle 3, participants did not receive any AI suggestion, i.e. control group), coherent suggestion (after puzzle 3, participants received a suggestion consistent with prior frames, i.e. 4 12 16 28 44), incoherent suggestion (after puzzle 3, participants received the same suggestion as in Study 1, which is inconsistent with prior frames, i.e. 4 12 40 28 36), and absurd suggestion (after puzzle 3, participants received a suggestion inconsistent with prior frames and also incorrect, i.e. 4 12 48 28 36). The "coherent suggestion" was selected to replicate the pattern of puzzle 2 (see Fig. 1), whereas the "absurd suggestion" was obtained by editing the "incoherent suggestion" already provided in Study 1 and Study 2. The (in)coherence of each suggestion with prior frames is determined by whether it is aligned with the task-related frame (i.e. the pattern is a monotone increasing function) that we primed in the first three puzzles.

5.2.1. Rewarding

The rewarding system was the same as in Study 1. In this case, the fixed reward was $2.36, while the variable component was $1.77. On average, each participant received a total reward of $3.91 ($16.64 per hour).

5.3. Measures

5.3.1. Dependent variables

This study relies on four dependent variables. The AI suggestion acceptance, augmentation type and completion time were operationalized as in Study 1.

Finally, performance is operationalized as answer correctness, that is a dichotomous variable taking a value of 1 if the participants' answer to the fourth puzzle coincided with a possible pattern. Wrong answers, which encompassed both instances of calculation errors leading to the identification of implausible patterns, and reliance on the absurd AI suggestion, took a value of 0.

5.3.2. Independent variable

The AI suggestion type is the independent variable of this study. It is a nominal variable that, depending on the experimental group, takes the value of 1 for the "no suggestion" group, 2 for the "coherent suggestion" group, 3 for the "incoherent suggestion" group, and 4 for the "absurd suggestion" group.

5.3.3. Control variables

The same control variables as in Study 1 were included, with the addition of further variables to account for individual cognitive differences that could influence the results. In particular, we controlled for the following variables. Cognitive reflection has been measured through three items from a validated translation of the CRT scale (Frederick, 2005). Risk aversion has been measured using the corresponding dimension from the propensity to trust scale of Ashleigh et al. (2012). Following previous work (Keding and Meissner, 2021), we have assessed trust in AI by adapting the items of Al-Natour et al. (2011) to this specific technology. The items were measured on a 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). Participants indicated the extent to which they agreed with the following statements: *"In general, I believe that artificial intelligence is competent." "In general, I believe that artificial intelligence is benevolent." "In general, I believe that artificial intelligence has high integrity." "Overall, I believe that artificial intelligence is reliable."*

5.4. Results

As in study 1, we present the results without control variables, highlighting in the text the main differences when those are included.

The descriptive statistics and intercorrelations among variables are reported in Table 4.

The number of correct answers was highest in the "incoherent suggestion" group (100 %), followed by the "coherent suggestion" group (90 %), the "no suggestion" group (89.7 %), and, finally, the absurd group (65 %). We adopted a binary logistic regression to examine the effect of suggestion type on the odds of providing a correct answer. The logit model correctly classified 86.4 % of the cases. Its fit with the data significantly improved compared to the null model (χ^2 (1) = 25.120, $p <$ .001, Nagelkerke R^2 = 0.262). The type of AI suggestion was significantly related to the odds of providing a correct answer to the last puzzle (Wald = 9.828, p = .020). When comparing the experimental groups, the only group that significantly differed from the "no suggestion" group in the odds of providing a correct answer was the absurd suggestion group ($\beta = -1.550$, OR = 0.212, 95 % CI [0.063, 0.720], Wald = 6.185, p = .013). Specifically, they had 4.7 times lower odds of giving a correct answer than the reference group.

Regarding completion time, participants exposed to the absurd suggestion showed the highest mean completion time (117.89 s), followed by those in the "incoherent suggestion" group (107.49 s), the "coherent suggestion" group (88.04 s), and the "no suggestion" group (74.77 s).

Our findings show a marginal significance (F(1,160) = 3.59, p = .06, η^2 = 0.022) of the positive relationship between having received an AI suggestion and the overall completion time of puzzle four. Comparisons of completion time across the four experimental conditions, using the "no suggestion" group as the reference, indicated that there were no significant differences between this and the "coherent suggestion" group (t(158) = 0.691, p = .491, 95 % CI [−24.69, 51.24]). By contrast, the

Table 4
Descriptive statistics and correlations among Study 3 variables.

		M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	Completion time	191.68	116.5																										
2	Algorithmic frame stating time	14.78	11.00	0.30*																									
3	Frame stating time	10.4	7.43	0.37**	0.11																								
4	Frame assuming time	1.88	4.04	0.31*	-0.11	0.2																							
5	Direction setting time	3.38	5.74	0.30*	-0.07	0.30*	0.16																						
6	Algorithmic direction setting time	6.02	8.00	0.44***	0.12	0.27	-0.05	0.05																					
7	Evaluation time	13.44	27.1	0.27	-0.12	0.15	0.32*	0.2	0.00																				
8	Decision time	13.34	8.66	0.16	-0.02	-0.09	-0.16	0.30*	0.18	0.14																			
9	Algorithmic search time	73.04	76.99	0.83***	0.41**	0.1	0.1	-0.02	0.44***	-0.02	-0.03																		
10	Search time	28.36	36.03	0.54***	-0.07	0.40**	0.36*	0.39**	0.12	0.07	0.18	0.12																	
11	Search evaluation time	27.04	21.67	0.54***	-0.11	0.19	0.16	0.27	0.03	0.04	0.02	0.32*	0.36*																
12	Suggestion acceptance (0/1)	0.22	0.42	0.08	0.05	-0.11	0.33*	-0.18	0.13	0.17	0.01	0.19	-0.29*	-0.01															
13	Unengaged refusal (0/1)	0.26	0.44	-0.56***	-0.42**	-0.35*	-0.18	-0.1	-0.34*	-0.21	-0.2	-0.56***	-0.07	-0.04	-0.31*														
14	Unengaged acceptance (0/1)	0.12	0.33	0.02	0.11	-0.14	0.13	-0.05	0.25	-0.04	0.09	0.15	-0.18	-0.25	0.70***	-													
15	Engaged refusal (0/1)	0.52	0.51	0.42**	0.33*	0.40**	-0.12	0.24	0.19	0.05	0.17	0.34*	0.30*	0.04	-0.55***	-	-												
16	Engaged acceptance (0/1)	0.1	0.3	0.09	-0.05	-0.01	0.31*	-0.2	-0.09	0.27	-0.09	0.09	-0.2	0.26	0.63***	-	-	-											
17	Unexpected failure of prior frames (0/1)	0.5	0.51	0.33*	0.27	0.14	-0.11	0.16	0.17	0.01	0.1	0.43**	-0.05	0.04	0.14	-	-	-	-										
18	BFExtraversion	5.24	1.16	0.02	-0.12	0.09	0.09	-0.01	-0.26	0.08	-0.2	-0.08	0.14	0.21	0.04	0.09	-0.08	-0.11	0.13	-0.09									
19	BFAgreeableness	5.36	1.13	0.19	0.04	-0.07	0.04	-0.1	0.03	0.31*	0.00	0.28	-0.18	-0.01	0.09	-0.09	-0.06	0.00	0.19	0.16	-0.21								
20	BFConscientiousness	5.82	1.13	0.23	-0.21	0.1	0.14	-0.04	-0.07	0.12	-0.25	0.24	0.13	0.2	-0.04	0.12	-0.13	-0.07	0.08	-0.09	0.57***	0.22							
21	BFEmotionalstab	4.12	1.54	-0.05	-0.08	-0.06	0.16	0.16	-0.12	0.27	0.09	-0.22	0.03	0.12	-0.04	-0.09	-0.09	0.12	0.04	-0.05	-0.04	-0.08	-0.18						
22	BFOpenness	5.19	1.11	-0.11	-0.03	-0.09	-0.31*	-0.12	-0.01	0.04	0.22	-0.08	-0.16	-0.06	0.04	0.08	0.05	-0.11	0.00	-0.12	0.24	0.07	0.05	-0.2					
23	Propensity to trust Technology	3.68	0.71	-0.25	-0.30*	0.12	-0.07	-0.18	0.09	0.04	-0.17	-0.20	-0.28	0.00	0.13	-0.1	-0.01	-0.02	0.18	-0.06	0.14	0.18	0.06	0.15	0.33*				
24	Age	22.8	2.49	-0.08	-0.11	-0.03	-0.16	0.23	-0.1	-0.02	0.17	-0.14	-0.03	0.13	-0.23	0.01	-0.17	0.18	-0.14	-0.1	0.01	-0.14	-0.03	0.16	-0.08	0.01			
25	Gender (male = 1, female = 2)	1.68	0.47	0.12	-0.13	0.03	-0.01	0.13	0.05	0.13	0.14	0.12	0.2	0.02	-0.05	0.11	0.12	-0.06	-0.2	0.00	0.29*	0.24	0.37**	-0.54***	0.22	-0.05	0.12		
26	Task skill	3.24	0.8	-0.14	-0.01	-0.09	-0.18	-0.13	-0.1	0.05	0.04	-0.08	-0.26	0.06	-0.28*	0.11	-0.42**	0.14	0.07	-0.25	-0.23	-0.01	0.00	-0.01	0.05	-0.01	0.23	-0.12	

Note. Correlations between unengaged refusal, unengaged acceptance, engaged refusal, and engaged acceptance are not displayed because these variables represent mutually exclusive responses from a single-choice question. * p ≤ .05; ** p ≤ .01; *** p ≤ 0.001.

differences between the reference group and the "absurd suggestion" and "incoherent suggestion" conditions were found to be significant (t(144) = 2.24, p = .026, 95 % CI [5.16, 81.08]) and marginally significant (t(158) = 1.73, p = .085, 95 % CI [−4.58, 70.02]), respectively. When control variables are included, the difference for the "incoherent suggestion" condition also becomes significant (t(158) = 1.982, p = .049, 95 % CI [0.104, 77.822]). Overall, the groups receiving a suggestion that defied prior frames displayed an increase in completion time compared to the control group.

For what concerns AI suggestion acceptance, we adopted a binary logistic regression to examine the effect of suggestion type on the odds of accepting the AI suggestion. For this analysis, we relied on the subsample of participants who received an AI suggestion (n = 123). The logit model correctly classified 66.7 % of the cases. The improvement of the null model due to the fit was marginally significant (χ^2 (1) = 5.720, p = .057, Nagelkerke R^2 = 0.063). The relation between the type of suggestion and the propensity to conform to AI was marginally significant (Wald = 5.609, p = .061). In our regression, the "coherent suggestion" group was adopted as the reference group. The odds for the "absurd suggestion" group to conform to the AI suggestion were not significantly different from those of the reference group (β = −0.747, OR = 0.474, 95 % CI [0.189, 1.186], Wald = 2.546, p = .111), while the "incoherent suggestion" group showed a significantly lower propensity to accept the suggestion compared to the reference group (β = −1.094, OR = 0.335, 95 % CI [0.131, 0.858], Wald = 5.190, p = .023). These results indicate that, differently from the "incoherent suggestion" group, participants exposed to the absurd suggestion did not differ significantly in terms of suggestion acceptance from the group showing the highest level of conformity. This outcome may be explained by the fact that, for the "incoherent suggestion" group, plausible explanations for the suggestions could be generated and evaluated against alternative solutions. By contrast, the absurd suggestion left room only for blind acceptance or outright rejection.

To evaluate this explanation, we conducted an additional binary logistic regression to test how the type of suggestion affected the odds of displaying an "unengaged acceptance" augmentation type. The overall model did not significantly improve the predictive accuracy compared to the null model (χ^2 (1) = 3.871, p = .144, Nagelkerke R^2 = 0.049). Overall, the model showed that suggestion type was not a significant predictor of the odds of displaying an "unengaged acceptance" augmentation type (Wald = 3.747, p = .154). When analyzing comparisons across groups, using the "coherent suggestion" group as the reference, no significant difference was found with the "incoherent suggestion" group (β = 0.470, OR = 1.600, 95 % CI [0.476, 5.374], Wald = 0.578, p = .447), while the "absurd suggestion" group displayed a positive and marginal difference (β = 1.099, OR = 3.000, 95 % CI [0.945, 9.528], Wald = 3.472, p = .062). This difference became significant when control variables were included in the model (β = 1.655, OR = 5.234, 95 % CI [1.249, 21.933], Wald = 5.127, p = .024), thus providing support for our explanation.

Finally, we analyze the impact of the type of suggestion on the odds of displaying augmentation patterns other than unengaged refusal through a multinomial logistic regression. The model shows a significant improvement over the null model ($\chi^2(6) = 33.16$, $p < .001$). The results indicate a significant increase in the odds of displaying engaged acceptance rather than unengaged refusal for the "coherent suggestion" group compared to the "absurd suggestion" group (β = 21.166, p < .001).[2]

[2] Note that the regression outcomes display extreme odds ratio and wide confidence intervals as no participants in the "absurd suggestion" group exhibited engaged acceptance.

6. General discussion

This work started with the research question "Under which conditions do individuals commit to different types of augmentation, and how does exposure to unexpected AI suggestions influence their reasoning?". In answering this question, we make multiple contributions to the literature on human-AI augmentation (Krakowski, 2025; Lebovitz et al., 2022; Daugherty and Wilson, 2018; Dwivedi et al., 2023).

First, by focusing on the subjective and situated conditions that influence individuals' reactions to unexpected AI suggestions, we highlight, coherently with sensemaking theory, the primary importance of the state of individuals' frames of reference.

Our findings show that when these frames remain valid and AI suggestions contrast with them, individuals are more likely to reconfirm their cognitive structures and interpret the AI suggestions as instances of malfunctioning. When these frames are unstable, for example because they have been challenged by unexpected failures as in our studies, individuals are more inclined to explore explanations for the opaque AI suggestions that they would have, otherwise, ignored (engaged patterns). Under such circumstances, when individuals fail in the sensemaking task, they may also display an increased likelihood of blindly conforming their judgment to the opaque AI suggestion (unengaged acceptance pattern). The disruption of prior frames produces these effects, as it was observed to generate puzzlement and hinder participants' self-confidence, a feeling that was managed by anchoring their reasoning to the AI suggestion. While prior literature focused on discomfort and confusion as outcomes of receiving unexpected AI suggestions (Lebovitz et al., 2022), our study also shows that these feelings are a requirement for incorporating unexpected information into the reasoning process. Finally, we present qualitative evidence on the individual factors that made participants more or less prone to anchor to AI after receiving unexpected negative feedback.

Second, we provide evidence of the effects that unexpected AI suggestions may produce. Human-AI augmentation also assumes that AI can provide answers that individuals do not anticipate, from which they may draw to achieve better decision outcomes. In line with existing literature (Lebovitz et al., 2022; Jussupow et al., 2021), our findings show that integrating unexpected AI suggestions into one's reasoning is an effortful and time-consuming process (van den Broek et al., 2021; Lebovitz et al., 2022). This challenges the notion of increased efficiency that is often associated with the adoption of AI technologies in organizations. Furthermore, we move beyond the time dimension by showing the impact of unexpected AI suggestions on performance. In particular, Study 3 shows that wrong suggestions generate a higher likelihood of blind reliance (unengaged acceptance) compared to other suggestion types, and lead to significant performance deterioration. Finally, the present study provides qualitative evidence on how unexpected suggestions affect human reasoning by examining participants' reasoning patterns through the think-aloud methodology. On the one hand, it offers an initial nuanced description of how each augmentation type unfolded. On the other hand, it highlights the need for more refined distinctions beyond the categories of engaged and unengaged augmentation in order to capture reactions to AI.

Finally, our findings also contribute to the literature on sensemaking (Jussupow et al., 2022; Weick, 1995; Maitlis and Christianson, 2014). The extant literature reveals that the unexpected disconfirmation of prior cognitive frames may trigger what Maitlis et al. (2013) define as integrative sensemaking, namely a specific type of sensemaking "characterized by a heightened sensitivity to whether new cues are consistent or inconsistent with the emerging account of a situation, such that accounts are continuously and critically evaluated with respect to their plausibility" (Maitlis et al., 2013, p. 230). As in our case, integrative sensemaking processes render individuals more open to attending to cues that would otherwise remain unnoticed and to elaborating accounts of events that deviate from pre-existing frames. According to the theory, this bottom-up approach should "lead to more precise constructions of a

situation based on more critical analyses of new information" (Maitlis et al., 2013, pp. 230–231). Our study suggests that this may not always be the case. If individuals, during integrative sensemaking processes, are confronted with opaque AI suggestions, they may be more likely to focus on constructing plausible explanations that justify the AI suggestion, disregarding the original problem and overlooking alternative solutions. We define this phenomenon as problematization pivoting.

Problematization pivoting differs from the well-known process of updating, which is "the process of revising provisional sensemaking to incorporate new cues" (Christianson, 2019, p.45). Similarly to updating, problematization pivoting occurs after sensemaking has already been initiated and an interruption, such as unexpected negative feedback, has been encountered. Yet, while updating involves refining existing accounts of a situation, problematization pivoting abandons the original problem focus of finding the correct solution and redefines the task around making sense of the suggestion of the AI. Our qualitative findings show that this is the case to the extent that participants who could not identify a plausible explanation for the answer of the AI did not attempt to search for alternative solutions; instead, they accepted the suggestion of the AI without question, despite the penalties this decision entailed. Problematization pivoting therefore alters not only the interpretive frames, in our task beliefs about what constitutes a pattern, but also the very definition of the problem, which shifts from finding a pattern to explaining the pattern proposed by the AI, under the assumption that it must be correct.

Overall, these findings challenge the ideas of augmentation and integrative sensemaking, showing how creating accounts to explain an AI diagnosis or an AI financial investment decision may not lead to more accurate representations of the situation, but merely to plausible narratives justifying hallucinations. AI does not provide hints, but final answers. Unlike human suggestions, which can be confirmed or disconfirmed through the analysis of the reasoning process, AI suggestions remain opaque, with their underlying logic closed to external scrutiny and open to interpretation that may confer them plausibility. This final point carries further implications for the literature on augmentation, showing that the additional time required by engaged augmentation may not necessarily lead to more accurate judgments.

6.1. Limitations and avenues of future research

This study has limitations which create opportunities for further research.

Our findings are based on laboratory studies in which participants solved numerical puzzles with AI suggestions. While this design allowed us to isolate cognitive mechanisms by minimizing confounding factors, it does not capture the complexity and ambiguity of organizational life, thereby limiting the generalizability of our findings. Strategic decisions in hiring, investment, and product development often allow for multiple plausible and creative interpretations, enabling decision-makers to reconcile unexpected AI suggestions with existing frames of reference. By contrast, the mathematical nature of our task hinders creativity and made such reconciliation impossible: participants could only treat the AI suggestion as an error if unwilling to question their existing frames. We hypothesize this likely increased resistance to AI, making our findings a conservative estimate of how humans might respond in real and more ambiguous organizational contexts. Future research should examine the extent to which our results hold in decision domains that more closely resemble organizational reality, and how contextual factors may foster or constrain the dynamics we identified. Moreover, our studies focused on socially distanced forms of reasoning (Maitlis et al., 2013). Even in such such circumstances, it is important to remark that sensemaking is a social accomplishment, as individual actors are constantly influenced by networks of beliefs, meanings, and social norms that affect their cognition and direct their behavior. Future research may study how the exposure to AI influences collective problem-solving processes. Finally, our findings highlight the need for further research to comprehend the relationship between sensemaking and AI technologies. For instance, considering that engaged forms of augmentation are more likely to occur after the disruption of existing frames, future research could investigate the role of sensebreaking practices (Pratt, 2000) in facilitating the effective incorporation of AI technology in organizational decision processes.

6.2. Practical implications

This study holds significant implications for various stakeholders, including policymakers and decision-makers. Our findings indicate that individuals are more likely to be influenced by AI suggestions following the occurrence of unexpected failures of their frames. These events are typically unprecedented or extremely rare, potentially coinciding with situations underrepresented in existing datasets, in which AI technologies may struggle too. In other words, our work highlights that individuals tend to be more influenced by AI technologies precisely when those technologies are more likely to produce incorrect results. These findings are particularly relevant in a broad range of AI applications, such as medical diagnoses, predictive maintenance, and self-piloted vehicles, where even a single incorrect action may lead to severe consequences (Oliver et al., 2017).

Further societal risks stem from the phenomenon of problematization pivoting. Our findings suggest that efforts to validate biased algorithmic decisions may influence individual behaviors, affecting the data generated for training AI technologies. This may transform algorithmic outputs into self-fulfilling prophecies. In contexts such as recruitment (van den Broek et al., 2021) or predictive policing (Waardenburg et al., 2022), this dynamic may risk reinforcing and perpetuating existing biases. Furthermore, we have seen how problematization pivoting also entails the lack of search for solutions independent from the AI suggestions. This shows that the presence of AI could prevent more creative search efforts to occur. In this context, policymakers and managers should prioritize training and educating employees on how to handle situations involving potentially misleading algorithmic suggestions. Additionally, they should encourage design practices (e.g., XAI practices) aimed at minimizing the occurrence of augmentation failures.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.techfore.2025.124491.

CRediT authorship contribution statement

Domenico di Prisco: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Silvia Dello Russo:** Writing – review & editing, Validation, Supervision, Software, Resources, Methodology, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4.0 in order to improve the readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding statement

This research received financial support from the Academia Italiana di Economia Aziendale (AIDEA).

Declaration of competing interest

No financial and personal relationships with other people or

organizations could inappropriately influence or bias the results of the study.

Acknowledgments

We wish to thank Luigi Marengo for the support and comments provided on different versions of the paper. We also thank Ginevra Assia Antonelli Gerardo Patriotta and Tamara Thuis for their feedback on early versions of this manuscript. We are also grateful to Cinzia Calluso, Daniella Laureiro-Martinez, Ekaterina Jussupow, Elena Bruni, Maria Giovanna Devetag and Sarah Lebovitz for their insightful comments. Preliminary drafts of this work were presented at: EGOS 2023 in the sub-theme "Explaining AI in the Context of Organizations", EURAM 2023 in the track "Artificial Intelligence and Digital Strategies" and in the Paradox & Plurality Summer School 2022. The authors wish to thank the participants at these conferences for their valuable feedback. Finally, they are grateful to AIDEA for the financial support provided for this research.

Data availability

Data of quantitative studies available under requests. Qualitative data not available because confidential.

References

Abolafia, M.Y., 2010. Narrative construction as sensemaking: how a central bank thinks. Organ. Stud. 31 (3), 349–367.
Aguinis, H., Villamor, I., Ramani, R.S., 2021. MTurk research: review and recommendations. J. Manag. 47 (4), 823–837.
Al-Natour, S., Benbasat, I., Cenfetelli, R., 2011. The adoption of online shopping assistants: perceived similarity as an antecedent to evaluative beliefs. Journal of the Association for Information Systems 12 (5), 347–374.
Anthony, C., 2021. When knowledge work and analytical technologies collide: the practices and consequences of black boxing algorithmic technologies. Adm. Sci. Q. 66 (4), 1173–1212.
Appio, F.P., Hernandez, C.T., Platania, F., Schiavone, F., 2025. AI narratives in fiction and media: exploring thematic parallels in public discourse. Technovation 142, 103201.
Ashleigh, M.J., Higgs, M., Dulewicz, V., 2012. A new propensity to trust scale and its relationship with individual well-being: implications for HRM policies and practices. Hum. Resour. Manag. J. 22 (4), 360–376.
Assis, A., Dantas, J., Andrade, E., 2025. The performance-interpretability trade-off: a comparative study of machine learning models. Journal of Reliable Intelligent Environments 11 (1), 1.
Baer, I., Waardenburg, L., Huysman, M., 2025. What is augmented? A metanarrative review of AI-based augmentation. Journal of the Association for Information Systems 26 (3), 760–798.
Balasubramanian, N., Ye, Y., Xu, M., 2022. Substituting human decision-making with machine learning: implications for organizational learning. Acad. Manage. Rev. 47 (3), 448–465.
Bauer, K., von Zahn, M., Hinz, O., 2023. Expl(AI)ned: the impact of explainable artificial intelligence on users' information processing. Information Systems Research 34 (4), 1582–1602.
Becker, T.E., Atinc, G., Breaugh, J.A., Carlson, K.D., Edwards, J.R., Spector, P.E., 2016. Statistical control in correlational studies: 10 essential recommendations for organizational researchers. J. Organ. Behav. 37 (2), 157–167.
Bolinger, M.T., Josefy, M.A., Stevenson, R., Hitt, M.A., 2022. Experiments in strategy research: a critical review and future research opportunities. Journal of Management 48 (1), 77–113.
Burrell, J., 2016. How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. 3 (1), 1–12.
Christianson, M.K., 2019. More and less effective updating: the role of trajectory management in making sense again. Adm. Sci. Q. 64 (1), 45–86.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.
Cornelissen, J.P., Werner, M.D., 2014. Putting framing in perspective: a review of framing and frame analysis across the management and organizational literature. Academy of Management Annals 8 (1), 181–235.
Cornelissen, J.P., Mantere, S., Vaara, E., 2014. The contraction of meaning: the combined effect of communication, emotions, and materiality on sensemaking in the Stockwell shooting. Journal of Management Studies 51 (5), 699–736.
Cruz, I.F., 2024. Expert-AI pairings: expert interventions in AI-powered decisions. Inf. Organ. 34 (4), 100527.
Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.
Daugherty, P., Wilson, H.J., 2018. Human + Machine: Reimagining Work in the Age of AI. Harvard Business Review Press, Boston, MA.
Davenport, T.H., Kirby, J., 2016a. Just how smart are smart machines? MIT Sloan Manag. Rev. 57 (3), 21.
Davenport, T.H., Kirby, J., 2016b. Only Humans Need Apply: Winners and Losers in the Age of Smart Machines. HarperCollins, New York.
Dwivedi, Y.K., Sharma, A., Rana, N.P., Giannakis, M., Goel, P., Dutot, V., 2023. Evolution of artificial intelligence research in technological forecasting and social change: research topics, trends, and future directions. Technological Forecasting and Social Change 192, 122579.
Ericsson, K.A., 2003. Valid and non-reactive verbalization of thoughts during performance of tasks: towards a solution to the central problems of introspection as a source of scientific data. J. Conscious. Stud. 10, 1–18.
Ericsson, K.A., Simon, H.A., 1980. Verbal reports as data. Psychol. Rev. 87, 215–251.
Evans, J.S.B., 2016. Reasoning, biases and dual processes: the lasting impact of Wason (1960). Q. J. Exp. Psychol. 69 (10), 2076–2092.
Fleischer, A., Mead, A.D., Huang, J., 2015. Inattentive responding in MTurk and other online samples. Ind. Organ. Psychol. 8 (2), 196–202.
Frederick, S., 2005. Cognitive reflection and decision making. J. Econ. Perspect. 19 (4), 25–42.
Fügener, A., Grahl, J., Gupta, A., Ketter, W., 2022. Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. Information Systems Research 33 (2), 678–696.
Glikson, E., Woolley, A.W., 2020. Human trust in artificial intelligence: review of empirical research. Academy of Management Annals 14 (2), 627–660.
Gosling, S.D., Rentfrow, P.J., Swann Jr., W.B., 2003. A very brief measure of the big-five personality domains. J. Res. Pers. 37 (6), 504–528.
Haque, A.B., Islam, A.N., Mikalef, P., 2023. Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change 186, 122120.
Jacovi, A., Marasović, A., Miller, T., Goldberg, Y., 2021. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), pp. 624–635.
Jussupow, E., Spohrer, K., Heinzl, A., Gawlitza, J., 2021. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Information Systems Research 32 (3), 713–735.
Jussupow, E., Spohrer, K., Heinzl, A., 2022. Radiologists' usage of diagnostic AI systems: the role of diagnostic self-efficacy for sensemaking from confirmation and disconfirmation. Bus. Inf. Syst. Eng. 64 (3), 293–309.
Keding, C., Meissner, P., 2021. Managerial overreliance on AI-augmented decision-making processes: how the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. Technological Forecasting and Social Change 171, 120970.
Krakowski, S., 2025. Human-AI agency in the age of generative AI. Information and Organization 35 (1), 100560.
Laureiro-Martinez, D., Arrieta, J.P., Brusoni, S., 2023. Microfoundations of problem solving: attentional engagement predicts problem-solving strategies. Organization Science 34 (6), 2207–2230.
Lebovitz, S., Lifshitz-Assaf, H., Levina, N., 2022. To engage or not to engage with AI for critical judgments: how professionals deal with opacity when using AI for medical diagnosis. Organization Science 33 (1), 126–148.
Logg, J.M., Minson, J.A., Moore, D.A., 2019. Algorithm appreciation: people prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. 151, 90–103.
Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., Stumpf, S., 2024. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Inf. Fusion 106, 102301.
Lu, T., Zhang, Y., 2025. 1+ 1> 2? Information, humans, and machines. Inf. Syst. Res. 36 (1), 394–418.
Mahmud, H., Islam, A.N., Ahmed, S.I., Smolander, K., 2022. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technological Forecasting and Social Change 175, 121390.
Maitlis, S., Christianson, M., 2014. Sensemaking in organizations: taking stock and moving forward. Acad. Manag. Ann. 8 (1), 57–125.
Maitlis, S., Sonenshein, S., 2010. Sensemaking in crisis and change: inspiration and insights from Weick (1988). J. Manag. Stud. 47 (3), 551–580.
Maitlis, S., Vogus, T.J., Lawrence, T.B., 2013. Sensemaking and emotion in organizations. Organ. Psychol. Rev. 3 (3), 222–247.
Martens, D., Hinns, J., Dams, C., Vergouwen, M., Evgeniou, T., 2025. Tell me a story! Narrative-driven XAI with large language models. Decis. Support. Syst. 191, 114402.
Mcknight, D.H., Carter, M., Thatcher, J.B., Clay, P.F., 2011. Trust in a specific technology: an investigation of its components and measures. ACM Trans. Manag. Inf. Syst. 2 (2), 1–25.
Mujtaba, D.F., Mahapatra, N.R., 2019. Ethical considerations in AI-based recruitment. In: Proceedings of the 2019 IEEE International Symposium on Technology and Society (ISTAS), pp. 1–7.
Oliver, N., Calvard, T., Potočnik, K., 2017. Cognition, technology, and organizational limits: lessons from the air France 447 disaster. Organ. Sci. 28 (4), 729–743.
Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., Agreda, S.G., 2022. To trust or not to trust? An assessment of trust in AI-based systems: concerns, ethics and contexts. Technological Forecasting and Social Change 181, 121763.
Pratt, M.G., 2000. The good, the bad, and the ambivalent: managing identification among Amway distributors. Adm. Sci. Q. 45 (3), 456–493.

Pumplun, L., Peters, F., Gawlitza, J.F., Buxmann, P., 2023. Bringing machine learning systems into clinical practice: a design science approach to explainable machine learning-based clinical decision support systems. J. Assoc. Inf. Syst. 24 (4), 953–979.
Raisch, S., Fomina, K., 2025. Hybrid problem-solving with large language models: a reply to "iterative alternative evaluation" and "an assemblage perspective". Acad. Manage. Rev. 50 (2), 482–484.
Raisch, S., Krakowski, S., 2021. Artificial intelligence and management: the automation-augmentation paradox. Acad. Manage. Rev. 46 (1), 192–210.
Renz, S., Kalimeris, J., Hofreiter, S., Spörrle, M., 2024. Me, myself and AI: how gender, personality and emotions determine willingness to use strong AI for self-improvement. Technological Forecasting and Social Change 209, 123760.
Schuetz, S., Kuai, L., Lacity, M.C., Steelman, Z., 2025. A qualitative systematic review of trust in technology. J. Inf. Technol. 40 (1), 55–76.
Shrestha, Y.R., Ben-Menahem, S.M., Von Krogh, G., 2019. Organizational decision-making structures in the age of artificial intelligence. Calif. Manage. Rev. 61 (4), 66–83.
Simon, H.A., 1957. Administrative Behavior, 2nd ed. Macmillan, New York.
Starbuck, W.H., Milliken, F.J., 1988. Executives' perceptual filters: What they notice and how they make sense. In: Hambrick, D.C. (Ed.), The Executive Effect: Concepts and Methods for Studying Top Managers. JAI Press, Greenwich, CT, pp. 35–65.
Strauss, A.L., Corbin, J.M., 1998. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. SAGE, Thousand Oaks, CA.
van den Broek, E., Sergeeva, A., Huysman, M., 2021. When the machine meets the expert: an ethnography of developing AI for hiring. MIS Q. 45 (3), 1557–1580.
Waardenburg, L., Huysman, M., Sergeeva, A.V., 2022. In the land of the blind, the one-eyed man is king: knowledge brokerage in the age of learning algorithms. Organization Science 33 (1), 59–82.
Wang, X., Zeng, D., Dai, H., Zhu, Y., 2020. Making the right business decision: forecasting the binary NPD strategy in Chinese automotive industry with machine learning methods. Technol. Forecast. Soc. Chang. 155, 120032.
Wason, P., 1960. On the failure to eliminate hypotheses in a conceptual task. Q. J. Exp. Psychol. 12 (3), 129–140.
Weick, K.E., 1979. The Social Psychology of Organizing. Addison-Wesley, Reading, MA.
Weick, K.E., 1988. Enacted sensemaking in crisis situations. Journal of Management Studies 25 (4), 305–317.
Weick, K.E., 1990. The vulnerable system: an analysis of the Tenerife air disaster. Journal of Management 16 (3), 571–593.
Weick, K.E., 1993. The collapse of sensemaking in organizations: the Mann gulch disaster. Adm. Sci. Q. 38 (4), 628–652.
Weick, K.E., 1995. Sensemaking in Organizations. SAGE, Thousand Oaks, CA.
Zhang, Y., Liao, Q.V., Bellamy, R.K., 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on fairness, Accountability, and Transparency, pp. 295–305.

Domenico di Prisco is Assistant Professor at IÉSEG School of Management. His research is situated at the intersection of organization theory, information systems, and organizational behavior, with a focus on the relationship between technology and cognition and its implications for organizations and society.

Silvia Dello Russo is Associate Professor of Organizational Behavior and Human Resource Management at LUISS University, Italy. Her research spans three main lines, also intertwined: developmental HR practices, work motivation along the life span, and the relationship between individuals and their social context at work. She has published, among others, in Journal of Organizational Behavior, Journal of Occupational Health Psychology, Journal of Vocational Behavior, Human Resource Management, and Journal of Occupational and Organizational Psychology.