The Metacognitive Demands and Opportunities of Generative AI

Lev Tankelevitch*
Microsoft Research
Cambridge, United Kingdom
lev.tankelevitch@microsoft.com

Viktor Kewenig*†
University College London
London, United Kingdom
ucjuvnk@ucl.ac.uk

Auste Simkute†
University of Edinburgh
Edinburgh, United Kingdom
a.simkute@sms.ed.ac.uk

Ava Elizabeth Scott†
University College London
London, United Kingdom
ava.scott.20@ucl.ac.uk

Advait Sarkar
Microsoft Research
Cambridge, United Kingdom
advait@microsoft.com

Abigail Sellen
Microsoft Research
Cambridge, United Kingdom
asellen@microsoft.com

Sean Rintel
Microsoft Research
Cambridge, United Kingdom
serintel@microsoft.com

ABSTRACT

Generative AI (GenAI) systems offer unprecedented opportunities for transforming professional and personal work, yet present challenges around prompting, evaluating and relying on outputs, and optimizing workflows. We argue that metacognition—the psychological ability to monitor and control one's thoughts and behavior—offers a valuable lens to understand and design for these usability challenges. Drawing on research in psychology and cognitive science, and recent GenAI user studies, we illustrate how GenAI systems impose metacognitive demands on users, requiring a high degree of metacognitive monitoring and control. We propose these demands could be addressed by integrating metacognitive support strategies into GenAI systems, and by designing GenAI systems to reduce their metacognitive demand by targeting explainability and customizability. Metacognition offers a coherent framework for understanding the usability challenges posed by GenAI, and provides novel research and design directions to advance human-AI interaction.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; *User centered design*; *Interaction design theory, concepts and paradigms*; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

Generative AI, Metacognition, Human-AI interaction, User Experience Design, System Usability

ACM Reference Format:
Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642902

1 INTRODUCTION

Generative artificial intelligence (GenAI) systems—using models, like Large Language Models (LLMs), that can generate artefacts by using extensive parameters and training data to model and sample from a feature space [23]—have the potential to transform personal and professional work. Their potential stems from a unique combination of *model flexibility* (in their input/output space), *generality* (in their applicability to a wide range of tasks), and *originality* (in their ability to generate novel content) [157]. However, these same properties also pose a challenge for designing GenAI systems to be human-centered [34]. User studies reveal a range of usability challenges around prompting [208], evaluating and relying on outputs [156], and deciding on an automation strategy: whether and how to integrate GenAI into workflows [14, 154].

Recent work has sought to characterize the unique properties of GenAI and their potential effects on users [156, 157], and to offer technical or design roadmaps for designing human-centered GenAI [34, 197]. However, there is not yet a coherent understanding of the usability challenges of GenAI, much less one grounded in a theory of human cognition. Indeed, recent work has called for foundational research to understand how people interact with GenAI and AI more broadly [99, 104]. Here, we argue that *metacognition*—the psychological ability to monitor and control one's own thought processes [4, 62, 131, 181]—offers a valuable and unexplored perspective to understand and design for the usability challenges of GenAI. Firstly, we suggest that current GenAI systems impose multiple metacognitive demands on users; understanding these demands

[*] Both authors contributed equally to this research.
[†] The work was done when the co-author was employed at Microsoft.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642902

can help interpret and probe the identified and potentially novel usability challenges. Secondly, we suggest that the perspective of metacognitive demands offers new research and design opportunities for human-AI interaction.

The metacognitive demands of working with GenAI systems parallel those of a manager delegating tasks to a team. A manager needs to clearly understand and formulate their goals, break down those goals into communicable tasks, confidently assess the quality of the team's output, and adjust plans accordingly along the way. Moreover, they need to decide whether, when, and how to even delegate tasks in the first place. Among others, these responsabilities involve the metacognitive *monitoring* and *control* of one's thought processes and behavior [4, 62, 128, 181].

Analogously, current GenAI systems often require verbalized prompting, demanding self-awareness of task goals, and decomposition of tasks into sub-tasks. System outputs then need to be evaluated, requiring well-adjusted confidence in one's evaluation and prompting abilities, and metacognitive flexibility to iterate on the prompting strategy as necessary. Alongside the local interactions with GenAI systems, the generality of GenAI poses another, higher-level metacognitive demand: the challenge of knowing whether and how to incorporate GenAI into workflows—i.e., one's 'automation strategy' (see also [154]). This demands self-awareness of GenAI's applicability to, and impact on, one's workflow; well-adjusted confidence in manual versus GenAI-supported task completion; and metacognitive flexibility to adapt one's workflows as needed. We posit that these metacognitive demands are induced by GenAI's model flexibility, generality, and originality.[1] In §3, we draw on metacognition research and recent user studies of GenAI to illustrate these metacognitive demands and offer new research directions to probe them further.

These demands can be addressed in at least two complementary ways. Firstly, given that metacognitive abilities can be taught [45, 52, 126], we can *improve users' metacognition* via metacognitive support strategies that can be integrated into GenAI systems. Evidence-based metacognitive support strategies include those that help users in their planning, self-evaluation, and self-management [159]. Recent HCI work has begun to pursue this direction [173, 204], albeit without explicitly grounding it in metacognition; we suggest that a metacognitive lens offers new research and design directions for augmenting GenAI system usability.

Secondly, we can *reduce the metacognitive demand* of GenAI systems by designing task-appropriate approaches to GenAI explainability and customizability. We suggest that explainability can help offload metacognitive processing from the user to the system, and that existing explainability approaches can be augmented by considering metacognition. Likewise, we suggest that a metacognitive perspective can provide insights on approaching the end-user customizability of GenAI systems. In §4, we draw on intervention studies to improve metacognition and studies of GenAI prototypes and human-AI interaction to explore research and design directions that can address the metacognitive demands of GenAI. Critically, we also highlight how GenAI's model flexibility, generality, and originality can serve as a design solution to these demands. Finally, we discuss the relationship between cognitive load and addressing metacognitive demands, offering ways to manage their balance. In summary, our work makes three distinct contributions:

(1) We conceptualize and ground the usability challenges of GenAI in an understanding of human metacognition, drawing on research from psychological and cognitive science and recent GenAI user studies.
(2) We draw from research on metacognitive interventions, GenAI prototypes, and human-AI interaction to propose two directions for addressing the metacognitive demands of GenAI: improving users' metacognition, and reducing the metacognitive demands of GenAI.
(3) We use the metacognition lens to identify the need—and concrete directions—for further research into the metacognitive demands of GenAI, and design opportunities that leverage the unique properties of GenAI to augment system usability.

In the next sections, we define metacognition, summarizing key research findings (§2); illustrate the metacognitive demands of GenAI, focusing on prompting, evaluating and relying on outputs, and deciding on one's automation strategy (§3); and propose ways to address these metacognitive demands (§4).

2 WHAT IS METACOGNITION?

Metacognition as a concept was first popularized by developmental psychologist John H. Flavell in the late 1970s [64], as he tried to understand how children come to be aware of their own cognitive processes. Subsequently, Nelson and Narens [128] showed that while adults are able to reflect on their thoughts, they often fail to be aware of the premises underlying their decision-making, and do not analyze, understand, and control their thought processes objectively. Their 'metacognitive model' first distinguished between object-level and meta-level cognition. *Object*-level processes reflect the basic cognitive work of perceiving, remembering, classifying, deciding, and so on. *Meta*-level processes monitor those object-level processes to assess their functioning (e.g., assessing how well one grasped the gist of a text) and allocate resources appropriately (e.g., deciding to re-read the text). Since then, a growing line of research has linked improved metacognition to a range of benefits across different domains. Studies have shown that improved metacognition helps individuals with management of time, focus, and effort [212], problem-solving [67], academic performance [45, 52, 106, 126, 180, 214], emotional well-being [200], and overall decision-making [206].

As we argue in §3, alongside the promises of GenAI to transform work, it also poses usability challenges that can be fruitfully understood via metacognition. Nevertheless, the field of human-computer interaction (HCI) has so far considered metacognition

[1] Although the perspective of metacognition is equally relevant for understanding the usability of search engines and similar technologies, we focus our scope to GenAI systems as they are relevantly distinct from that of search engines [157]. Firstly, they are more flexible in their responsiveness to user prompts, in the range of implicit and explicit parameters available to users in their prompts, and in the multi-modality of their input/output space. Secondly, unlike search engines, they function as general-purpose tools, able to perform content generation, discrimination, and editing, among other functions (rather than merely retrieve existing content). Finally, unlike search engines, current GenAI systems are non-deterministic in their responses. As we aim to demonstrate here, all of these features place unique demands on users' metacognition and inform the design space of solutions to address these demands, a design space which necessarily extends beyond that of current search engines. Relatedly, we also note that Russell [50] proposed a connected idea of 'meta-literacy' for search engine usability (see also [112]); however, this work does not delve into the psychological and cognitive science of metacognition that is central to the current work.

mainly in the context of computer science education [107, 142]. The relative absence of metacognition research from many areas of HCI is surprising, considering that the early work on graphical user interfaces was, as Alan Kay concluded, *"solidly intertwined with learning"* [88]. One possible reason for this absence is the confusing plethora of existing and overlapping frameworks and theories on metacognition. From education [106] to management [90], healthcare [36], and even sports [111], many research disciplines have carved out their own approach to metacognition, producing multiple inconsistent terminologies and frameworks (for reviews see [131, 181]).

To structure our analysis of the metacognitive demands of GenAI systems, in §2.1-2.4 we present a simplified descriptive framework of metacognition, also summarized in Figure 1. In line with most common prior frameworks, we distinguish between metacognitive *knowledge* and *experiences*, two different sources of information for understanding one's own cognition [55, 131, 181], and between the metacognitive abilities of *monitoring* and *control*, through which one can assess and guide their own cognition [4, 62, 129].

2.1 Metacognitive knowledge and experiences

Metacognitive knowledge, being explicit, includes people's conscious understanding of aspects like their strategies (e.g., memory strategies [128]), reasoning abilities, decision-making, and beliefs [170].

Metacognitive experiences include anything that people can directly experience, and can be implicit, occurring without our direct intention or awareness [55]. This includes subjective feelings, like a feeling of familiarity, or the feeling that one has misunderstood a passage while reading, as well as other implicit cues that provide information about cognitive processing (e.g., 'processing fluency' cues, such as the speed at which a memory is retrieved) [4, 131].

Metacognitive knowledge and experiences are interrelated [64]. Metacognitive experiences can contribute to metacognitive knowledge—e.g., when feelings of difficulty during problem-solving become encoded as knowledge that one is poor at problem-solving. Metacognitive knowledge can also be retrieved during metacognitive experiences, for example, when one remembers that they are poor at problem-solving when experiencing a feeling of difficulty.

2.2 Metacognitive abilities: monitoring and control

Monitoring abilities involve the assessment of one's own thinking, whereas *control* abilities are those that directly guide one's own thinking. Our focus is on the monitoring and control abilities that are most relevant to concrete task-oriented metacognitive demands posed by GenAI (see [181] for a more in-depth taxonomy).

Relevant monitoring abilities for working with GenAI include self-awareness and adjustment of confidence. *Self-awareness* is the capacity to recognize one's own thoughts, emotions, and actions, as well as how these factors influence cognition [73, 212]. This includes having a clear awareness of one's specific goals and intentions—for example, *"What am I trying to convey with this email?"*. This ability is important for prompting GenAI and determining one's automation strategy (§3.1 and §3.3).

Confidence is one's self-assessment of one's cognitive abilities and their application to tasks [206]—e.g., *"How confident am I that I can write this email with the appropriate tone and level of detail?"* A 'well-adjusted' confidence distinguishes objectively correct and incorrect performance, and accurately matches one's abilities.[2] Confidence and its adjustment are central to decision-making and reasoning, especially in many aspects of human-AI interaction [4, 171] (§3.1, §3.2, and §3.3).

Relevant control abilities for working with GenAI include metacognitive flexibility and task decomposition. *Metacognitive flexibility* is the ability to adaptively shift cognitive strategies when encountering new information, when realizing that a current strategy isn't effective, or when the demands of the task change [32]—e.g., *"I recognize that my formal tone in my emails does not match the more conversational style of my new co-workers. I should therefore adjust my approach while still maintaining professionalism"*. It is a hallmark of creative problem-solving [144] and has been deemed essential for organizing and integrating a rapidly changing body of information [120]. Metacognitive flexibility is especially important when prompting and evaluating the output of GenAI (§3.1 and §3.2) and determining one's automation strategy (§3.3).

Task decomposition involves breaking down a task into concrete, actionable sub-tasks or steps. For instance, before writing an email, one might set clear objectives for what specific points to communicate—e.g., *"I want to clearly explain the status of the project and ask for feedback."* Then, one might decide on a structure for the email, laying out the most important aspects first. These abilities are especially important for prompting GenAI (Section 3.1). As the example suggests, task decomposition is not solely metacognitive because it often involves object-level cognitive processes.

Monitoring and control are interrelated [93, 95]. Monitoring (i.e., assessing our performance) affects control (e.g., by influencing a change in strategies). Control (e.g., changing strategies) can also provide feedback which affects monitoring (e.g., by altering the assessment of our performance).

2.3 Interrelationship between knowledge and experiences, and monitoring and control

Metacognitive knowledge and experiences interact with monitoring and, in turn, with control [181]. For example, adequate metacognitive knowledge of the strengths and weaknesses of one's strategies can affect the adjustment of confidence in a solution to a problem (i.e., metacognitive monitoring). Conversely, improving monitoring, such as by practicing self-awareness, can increase one's awareness of metacognitive experiences and knowledge. Likewise, metacognitive experiences can influence, and be influenced by, our metacognitive monitoring and control. For example, after experiencing a sense of misunderstanding, we might unconsciously adjust our sense of confidence (monitoring), and be prompted to re-read a passage (control). Similarly, the impact of metacognitive experiences might vary based on monitoring abilities. For example, a person with better monitoring might be more attuned to these

[2] While beyond the scope of this work, metacognition research distinguishes between two formal and independent aspects of confidence: *resolution* (also known as sensitivity), the ability for confidence judgments to distinguish correct and incorrect performance, and *calibration* (also known as bias), the extent to which confidence tends to be overall higher or lower than objective performance [65, 66]. We indicate this distinction in relevant points, but direct interested readers to the cited work for more information.

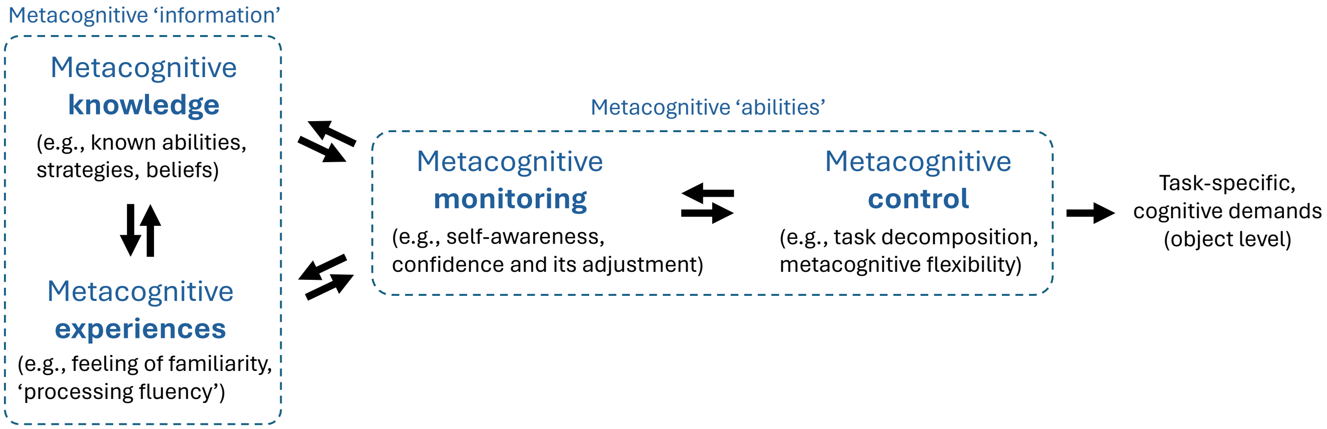

Metacognitive 'information'
Metacognitive knowledge
(e.g., known abilities, strategies, beliefs)
Metacognitive experiences
(e.g., feeling of familiarity, 'processing fluency')
Metacognitive 'abilities'
Metacognitive monitoring
(e.g., self-awareness, confidence and its adjustment)
Metacognitive control
(e.g., task decomposition, metacognitive flexibility)
Task-specific, cognitive demands (object level)

Figure 1: A simplified descriptive framework for metacognition. Metacognitive *knowledge* is the explicit understanding of one's abilities, strategies, and beliefs. Metacognitive *experiences* include things that people can directly experience, such as a feeling of familiarity or other implicit cues that provide information about cognitive processes. Metacognitive knowledge and experiences are interrelated in that experiences can become encoded as knowledge, and knowledge can be retrieved during experiences (§2.3). Both of these can influence (and be influenced by) metacognitive *monitoring*, which includes self-awareness, and confidence and its adjustment. Metacognitive monitoring, in turn, influences (and is influenced by) metacognitive *control* processes, such as metacognitive flexibility and task decomposition. Metacognitive control acts upon the (object-level) cognitive processes involved in a task. Arrows indicate directions of influence (§2.3).

experiences (thereby also making them less implicit) [212]. Thus, there is a tight interrelationship between metacognitive knowledge and experiences, and monitoring and control.[3]

2.4 Domain generality and specificity of metacognition

Whether metacognitive abilities, knowledge, and experiences are *domain-general* (universally applicable across different areas of knowledge, skills, or problem-solving) or *domain-specific* (pertain to a particular area of expertise to which they are finely tuned, such as math) is a matter of active debate, although there is evidence for both views [11, 44, 68, 118, 151]. What kind of metacognition a situation demands is likely context-dependent [44, 55], a particularly relevant consideration for GenAI given its generality across domains [157].

2.5 Heuristics and priming metacognition

People often implicitly and unintentionally rely on heuristics to guide their metacognitive monitoring and control [3]. For example, people guide their metacognitive control by implicitly relying on the ease of information processing ('processing fluency'). Information that is easily or fluently processed (e.g., in terms of reading) triggers less further cognitive processing relative to information that is more difficult to process [3]. Because processing fluency is subjectively experienced rather than consciously known, and because these heuristics are often activated implicitly, their activation represents a metacognitive experience [3, 187].[4] Manipulating the activation of these heuristics (e.g., via priming) can be used to improve metacognitive control: increasing subjective processing difficulty (e.g., by using a degraded font) stimulates more metacognitive control and thereby improves participants' performance on reasoning tasks that benefit from a more analytic processing style [7]. In §3, we discuss how engaging these heuristics when interacting with GenAI may influence users.

2.6 Improving metacognition

Metacognitive abilities can be taught and improved [45, 52]. Metacognitive interventions, such as training through feedback on metacognitive performance [30], can, for example, increase judgment accuracy—the ability to distinguish between one's own correct and incorrect metacognitive processing. Other interventions include providing feedback to adjust a person's mental model for a specific task [195], and using guided reflection to improve the ability to discern the reliability of outputs [180]. In §4.1, we discuss how some of these interventions can be used in practice to meet the metacognitive demands posed by GenAI.

2.7 Measuring metacognition

Metacognition researchers have devised a range of methods to measure different metacognitive abilities, both prospectively and retrospectively. Table 1 summarizes key methods from metacognition research relevant to exploring interactions with GenAI.

[3] Some metacognition theories view metacognitive knowledge and experiences not as separate from monitoring, but rather as *instances* of monitoring that can be either knowledge- or experience- based [4, 93, 128]. However, we distinguish between the two sets of concepts to emphasize the difference between the *ability* to monitor and control cognition, and *information* about cognition arising from metacognitive knowledge and experiences—e.g., the difference between the ability to be self-aware about one's memory (monitoring) and the information conveyed by a feeling of familiarity (an experience).

[4] In contrast, the conscious *knowledge* of these heuristics exemplifies metacognitive knowledge.

Table 1: Overview of some prospective and retrospective methods for exploring relevant metacognitive abilities. Using these methods to measure the metacognitive demands posed by GenAI and applying them for improving GenAI usability are promising opportunities for future research (see Table 2 and Table 3).

Ability	Type	Measure	Description
Self-awareness	Prospective	Think-aloud [59]	Users verbalize their thought process during a task.
	Prospective	Self-report [72]	Users report their perceived strengths and weaknesses.
	Prospective	Prediction log [56]	Users predict performance and feelings for an upcoming task.
	Retrospective	Reflective essay [76]	Users describe their thought processes after task completion.
	Retrospective	Interview [98]	Interviews focus on users' self-perception during or after a task.
	Retrospective	Assessment rubric [9]	Users assess their performance using a predefined rubric.
Confidence	Prospective	Judgment of learning (JOL) [129]	Users predict their performance before a test.
	Prospective	Self-rating [13]	Users rate their confidence in specific skills before a trial or task. Correlation between confidence and objective accuracy can estimate confidence *calibration* [127, 202]; *Meta-d'* is a derived metric capturing users' prospective confidence *resolution* independent of their calibration [66, 115] (see also [65, 145]).
	Prospective	Likelihood estimate [206]	Users estimate the likelihood of success in a future event.
	Retrospective	Self-Rating [24]	Users rate their confidence in their performance after a trial or task.
	Retrospective	Reflective journals [125]	Users reflect and comment on how confident they felt during the task.
Task decomposition	Prospective	Expectancy questionnaire [54]	Users set specific goals and plans before a task.
	Prospective	Self-regulated learning (SRL) microanalysis [37]	Users respond to prompts assessing strategy use and motivational beliefs.
	Prospective	Goal-setting worksheet [213]	Users fill out a survey about a task's value and expected success.
	Retrospective	Performance reviews [162]	Users evaluate their self-regulation strategies used in a task.
	Retrospective	Behavioural observations [140]	Recorded task sessions are coded for indicators of self-regulated learning.
	Retrospective	Reflective interview [98]	Interviews explore users' strategic planning, monitoring, and evaluation.
Metacognitive flexibility	Prospective	Cognitive flexibility scale [117]	Users describe how they solve problems in different contexts.
	Prospective	Task switching [124]	Users are tested on their ability to switch between task sets.
	Prospective	Category fluency [186]	Users list examples within categories in a given time.
	Retrospective	Post-task debrief [168]	Interviews about users' different strategies used and adaptability during the task.
	Retrospective	Solution review [86]	Users review and discuss the solutions they generated for a task.
	Retrospective	Error analysis [31]	Mistakes made during task performance are analyzed to understand metacognitive flexibility.

3 THE METACOGNITIVE DEMANDS OF GENERATIVE AI

As Sarkar et al. [156] notes, programming with GenAI may have *"far-reaching impact on [programmers'] attitudes and practices of authoring, information foraging, debugging, refactoring, testing, documentation, code maintenance, learning, and more"*. Other domains, such as design [70], writing [132], and data science [74] are likely to be experiencing similar changes with GenAI. We suggest that a core dimension underlying these changes is a greater demand on users' metacognition that is imposed when users have to (a) prompt GenAI systems, (b) evaluate and decide to rely on GenAI output, and (c) decide on one's workflow automation strategy: whether they should automate certain tasks with GenAI and how to automate them most effectively (previously described as *"critical integration"* [154]; see also [149]).

It is important here to distinguish between *metacognitive demand*—the need for extensive metacognitive monitoring and control for a task—and *cognitive load*, the total amount of mental effort required for a task [178]. Metacognitive demand contributes to cognitive load, but so do other aspects related to cognitive processing at the object (non-meta) level (i.e., metacognitive demand is sufficient but not necessary for increasing cognitive load). For example, as we illustrate below, prompting in current GenAI systems imposes a high metacognitive demand due to the need for self-awareness of goals, increasing cognitive load, while the interaction method of typing (rather than speaking) further increases cognitive load, albeit without much associated metacognitive demand. The relationship between metacognitive demand and cognitive load becomes relevant when considering interventions to support users' metacognition (see §4.3).[5]

This section covers each aspect of working with GenAI systems, describing how GenAI imposes high metacognitive monitoring and control demands on users (summarized in Figure 2). Not all GenAI systems impose the same type and extent of metacognitive demands due to differences in interface design and interaction modes; where relevant, we point out the implications of this. Throughout, we make concrete suggestions on future research to better understand these demands (summarized in Table 2).

3.1 Prompting generative AI systems

End-user studies suggest that prompting is challenging, with non-expert users making various errors and adopting ineffective strategies—a reflection of the demand on users' metacognitive monitoring and control [33, 41, 43, 85, 103, 174, 205]. During *prompt formulation*, the open-endedness of many current prompting interfaces requires users to have self-awareness of their specific task goals, and be able to decompose their tasks into smaller sub-tasks so as to verbalize these as effective prompts (Figure 2a). Next, iterative output evaluation and adjustment (*prompt iteration*) depends on users' confidence in their prompting ability, and metacognitive flexibility to adapt their prompting strategy (Figure 2b). We posit that these demands are exacerbated by GenAI's non-determinism and model flexibility (not to be confused with metacognitive flexibility) in terms of (a) the wide range of explicit and implicit parameters that users can adjust, and (b) systems' ability to work with prompts at a wide range of abstraction [156, 157].[6]

3.1.1 Prompt formulation: self-awareness and task decomposition.

In manual task completion, many implicit goals and intentions embedded within tasks can remain so without ever being verbalized. For example, when writing an email to a senior colleague, one might implicitly know to adopt a certain tone. Many GenAI systems require specification that the email is to a senior colleague and needs an appropriate tone. Moreover, it often requires that a task be broken up into sub-tasks (*"combine my content", "condense into two paragraphs", "update the tone"*). This demand for self-awareness and task decomposition is exacerbated by a particular type of model flexibility in GenAI: today's systems afford many parameters for end-users to adjust; these can be formal parameters like the model temperature, or a range of unspecified parameters that can be adjusted through text prompting (e.g., the tone, level of detail, or structure of a piece of text). This model flexibility and control afforded to users requires knowing what one wants to achieve and convey that explicitly and effectively to the system. Recent user studies of GenAI systems illustrate these demands.

In [208], non-expert participants used an LLM-based tool to improve a chatbot through prompting. One of the challenges they experienced was a struggle getting started.[7] Zamfirescu-Pereira et al. [208] interpret this as a design-stage barrier in end-user programming, reflecting some version of the implicit question, *"I don't even know what I want the computer to do"*. The self-awareness and explicitness demanded by prompting is also observed in LLM-supported writing. Dang et al. [41] define and compare *diegetic* prompting (instructions conveyed implicitly when users input text for the system to modify) and *non-diegetic* prompting (explicit instructions to the system). The latter is experienced as far more challenging by users as it *"forces writers to shift from thinking about their narrative or argument to thinking about instructions to the system"* [41]. Similar difficulty with non-diegetic prompting was observed among novice programmers in AI-assisted coding [83], and manufacturing designers co-creating with GenAI, who struggled to *"think through the design problem in advance"* [70]. These difficulties were exacerbated by the many parameters available to users who grappled with understanding and using them effectively [70, 83, 204].

The difference between diegetic and non-diegetic prompting points to the broader question around GenAI system interfaces and interaction modes and what they imply for the user experience and for productivity. For example, whereas diegetic prompting is easier, it affords less control than the alternative, with preferences differing across users [41]. Moreover, user experience and productivity may not always go hand in hand. For example, systems with non-diegetic prompting (e.g., ChatGPT) may be more challenging and time-consuming, yet the explicitness they require may plausibly act as a forcing function that ultimately trains metacognitive self-awareness and task decomposition, leading to higher quality output, assuming users persevere [155].[8] Nevertheless, in §4.1 we suggest that there are more effective and user-friendly ways of supporting metacognition.

Apropos of training, a key difference between expert and non-expert programmers—and by extension, expert and non-expert prompt writers—is an explicit approach to considering task requirements [208]. Expert programmers have advanced metacognitive monitoring and control, in that they are able to identify their specific goals and decompose them into concrete tasks [60]. One developer in [103] described their strategy with LLM-supported coding as, *"be incredibly specific with the instructions and write them as precisely as I would for a stupid collaborator"*. Likewise, users in [14] who decomposed the programming task into "*microtasks*"—"*well-understood and well-defined jobs*"—were able to work effectively with Copilot (see also [150, 188]). Beyond coding, manufacturing designers who successfully learned to co-create with GenAI abstracted and explained the problem to themselves [70].[9]

[5] For an in-depth theoretical discussion of metacognition (or the overlapping concept of self-regulated learning) and cognitive load, see [46, 160, 163, 189].

[6] A popular workaround to the challenge of prompting is 'prompt libraries' with detailed, task-specific prompts (see [176] for an overview). While helpful, ready-made prompts will rarely suit one's context precisely—the devil remains in the details. Moreover, ready-made prompts still require metacognitive ability to apply appropriately.

[7] Support getting started is a key user request for GenAI explainability; see §4.2.1.

[8] The potential discrepancy between user experience and productivity is reminiscent of that found in education, where the cognitive effort of effective learning is experienced negatively by students, leading to a divergence between *perceptions* of the learning experience and objective learning outcomes.

[9] As the above quotes suggest, task decomposition can often mean crafting a prompt as a set of discrete instructions that a system can interpret all at once, but it can also

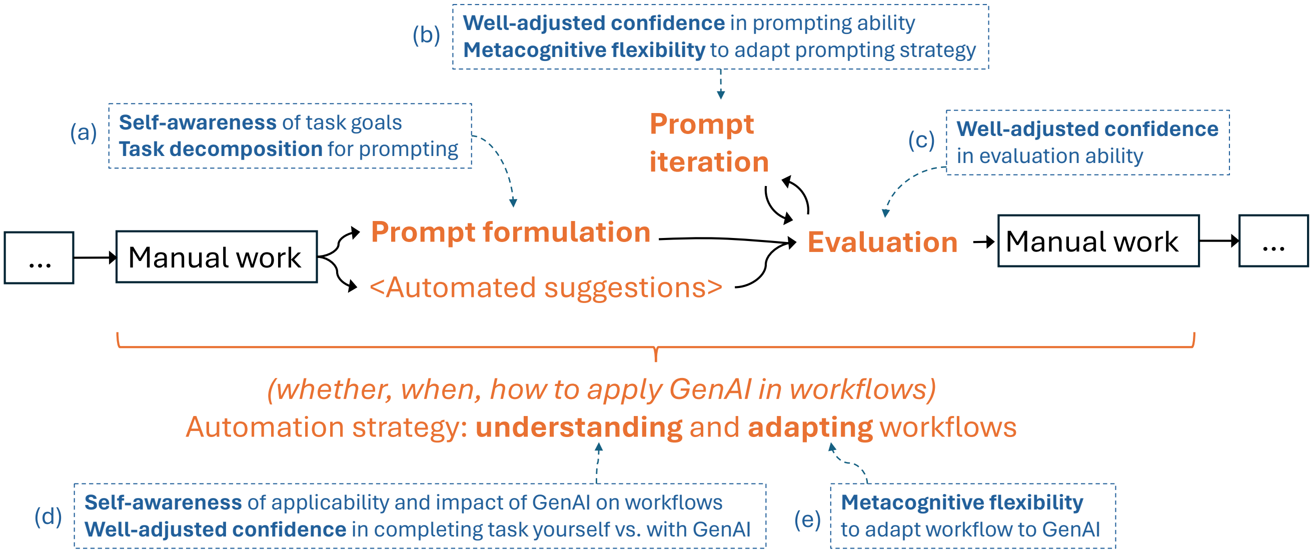

(b)
Well-adjusted confidence in prompting ability
Metacognitive flexibility to adapt prompting strategy
(a)
Self-awareness of task goals
Task decomposition for prompting
Prompt iteration
(c)
Well-adjusted confidence in evaluation ability
...
Manual work
Prompt formulation
<Automated suggestions>
Evaluation
Manual work
...
(whether, when, how to apply GenAI in workflows)
Automation strategy: understanding and adapting workflows
(d)
Self-awareness of applicability and impact of GenAI on workflows
Well-adjusted confidence in completing task yourself vs. with GenAI
(e)
Metacognitive flexibility to adapt workflow to GenAI

Figure 2: Metacognitive demands posed by generative AI at each point in a simplified user workflow. Often embedded within a workflow with manual tasks, users may first need to formulate a prompt, requiring metacognitive abilities including self-awareness of task goals and task decomposition (a). Systems that provide automated suggestions such as GitHub Copilot alleviate some of the demands associated with prompting. Depending on the output, iterating on the prompt may be necessary, which requires well-adjusted confidence in one's prompting ability and metacognitive flexiblity to adapt prompting strategies as necessary (b). Likewise, evaluating the output requires well-adjusted confidence in one's ability to judge its validity (c). Beyond the local interaction with a GenAI system, there is an overarching demand connected to understanding whether, when, and how to apply GenAI to one's workflows—one's 'automation strategy'. This requires self-awareness of how GenAI applies to and affects one's workflows, and well-adjusted confidence in the ability to complete tasks manually and with GenAI (d). Finally, it also requires metacognitive flexbility to adapt one's workflows as necessary (e).

We note that, although self-awareness and task decomposition are often required to some extent when interacting with GenAI systems, their pertinence increases as users concretize their usage intentions (e.g., achieving work or personal goals). For example, users interacting with GenAI systems for non-specific entertainment or exploration may worry less about their prompting strategy. At the same time, systems that support users' metacognition can help surface or clarify intentions originally hidden from users' self-awareness, thereby influencing their initial goals or lack thereof (see §4.1.2 for details). For example, users 'playing' with a system may be enabled to identify more concrete or diverse forms of play for them to explore. Thus, our framework of metacognitive demands is applicable to many use-cases. More generally, we do not assume that users' intentions and goals (or lack thereof) remain static during human-AI interaction, and, as per §4.1.2, suggest that systems can and should help users clarify their intentions and goals.

Future research should systematically examine how self-awareness and task decomposition ability moderate users' ability to control systems across interaction modes (e.g., diegetic vs. non-diegetic prompting), task contexts (e.g., creating a novel output vs. editing an existing artifact), and domains (e.g., writing vs. programming).

3.1.2 Prompt iteration: confidence adjustment and metacognitive flexibility.

After the initial prompt, the next common step is iteration: evaluating the output and adjusting the prompt accordingly (here we focus on prompting; see §3.2 on evaluating the output). Alongside maintaining awareness of their task goals, users need to (Figure 2b):

(a) evaluate the output with respect to their prompt,
(b) adjust their confidence in their prompting ability, to disentangle this from systems' capabilities (*"Is my prompt specific and clear enough; are system parameters set appropriately; is system performance generally poor on this task; or is this an 'unlucky' probabilistic output?"*),
(c) flexibly adjust their prompting strategy as needed (*"Should I adjust my prompt, adjust an earlier prompt, decompose my tasks into further sub-tasks, re-try with the same prompt…etc.?"*).

The range of possible explanations for a poor output makes confidence adjustment challenging, and the range of possible strategy adaptations demands high metacognitive flexibility from the user. This is exacerbated by the non-determinism of GenAI, particularly when tweaking one aspect of the prompt might unintentionally change a different aspect of the output [33]. This requires constantly maintaining awareness of one's task goals in the face of

require step-wise prompting, which can work sequentially to produce the desired output, or, in some cases, may require manual reassembly of multiple outputs.

ever-changing output, or risk getting derailed from the task by unexpected output, as some users were in [204].[10]

It is further exacerbated by a distinct type of model flexibility in GenAI systems (not to be confused with metacognitive flexibility): *"[generative AI systems] can generate plausible and correct results for statements at an extremely wide range of abstraction"*, which presents what Sarkar et al. [156] term a 'fuzzy abstraction matching' problem—it becomes difficult for users to discern a system's capabilities and to match one's intent and prompting accordingly (see also [61, 85] for similar conclusions).

Participants in [208] struggled with this (see also [43, 85, 205]). They were unable to choose the right prompting instructions, incorrectly expecting human capabilities; in some cases, underestimating the system's capabilities; and insisted on socially appropriate—rather than effective—ways of prompting. Zamfirescu-Pereira et al. interpret these challenges as stemming from *"over-generalization from limited experience"*, and *"a social lens that filtered participants' prompts… through expectations originating in human-human interactions"*. From a metacognitive perspective, over-generalization reflects poorly adjusted confidence; rather than maintaining an appropriately low confidence (about their own prompting), and gathering more evidence, participants drew confident conclusions based on limited evidence. Participants' insistence on a social lens may reflect a lack of self-awareness of their prompting approach, poorly adjusted confidence, and/or inflexibility in their strategies. To be clear, these challenges partly stem from a lack of feedback in the system about prompting effectiveness, leaving users to grapple with the fuzzy abstraction matching problem without support (see also §4.2.1 on explainability). However, that participants in [208] *"avoided effective prompt designs even after their interviewer encouraged their use and demonstrated their effectiveness"* suggests that this is also partly a metacognitive failure to notice and/or adjust their mental model of the system, signaling low metacognitive flexibility (see also [70, 204]).

Future research should systematically investigate how different aspects of GenAI systems—such as their non-determinism and model flexibility—impact users' ability to adjust their confidence in their prompting ability, flexibly adapt their prompting strategy, and update their mental model of these systems. For example, this could examine how the temperature setting of a model influences users' confidence and its adjustment, or how different levels of abstraction influence novice users' prompting strategies.

3.2 Evaluating and relying on generative AI outputs

Evaluating and relying on AI output requires users to maintain a well-adjusted confidence in their own domain expertise and ability to evaluate output (i.e., self-confidence; Figure 2c). The importance of this metacognitive demand is evidenced in recent research on AI-assisted decision-making, which finds that users' self-confidence is a key determinant of their reliance on AI responses, as discussed in §3.2.1 below. Confidence in the *system's* abilities is also important, and likely interacts with self-confidence, although here we focus on the latter as it is a metacognitive concept (i.e., an assessment of one's cognition via metacognitive monitoring).[11]

We posit that GenAI exacerbates the demand for a well-adjusted confidence in output evaluation. In this context, this includes confidence with 'good' *calibration*, meaning the overall confidence of a user in their output evaluation accurately matches objective performance; and with 'good' *resolution*, meaning the user's confidence can correctly discriminate a correct output.[12] The generative nature of GenAI—its 'originality' [157]—means that many user workflows will or have shifted from users *generating* content to *evaluating* it [149, 154], as already documented in programming [156] and manufacturing design [70]. Thus, users must maintain a well-adjusted level of confidence in their own ability to evaluate this output and not blindly accept generated content. Moreover, GenAI poses unique challenges to confidence adjustment that we discuss below.

3.2.1 AI output evaluation and reliance: confidence adjustment.

Recent work has investigated the role of self-confidence in evaluating and relying on AI output, although with discriminative models, rather than GenAI. For example, in AI-assisted decision-making in chess, participants' reliance on the AI was only significantly predicted by their self-confidence, and not by their confidence in the AI [35]. This dovetails with He et al. [77] who find that poor performers in a logical reasoning task tend to be overconfident—the Dunning-Kruger effect—leading to under-reliance on AI (see [182] for related findings). Lu and Yin [108] show that, in the absence of AI accuracy information, humans rely on their agreement with AI as a heuristic for their reliance on it, albeit only when they have a high self-confidence. Older human-automation interaction studies demonstrated a similar role for user self-confidence in influencing reliance on automation [47, 102, 114].

Analogous research in GenAI that explicitly measures and manipulates user self-confidence is missing, but user studies suggest a similar key role for well-adjusted confidence in output evaluation and reliance.[13] Programmers were reluctant to deeply review and repair AI-generated code, preferring instead to re-write the entire code themselves [188]. Similarly, manufacturing designers co-creating with GenAI were uncertain about how to interpret outputs and whether users or the system were responsible for addressing errors [70]. By contrast, programmers who were highly confident in their own ability actively questioned the AI code assistant when it produced confusing output [198].

The challenge of output evaluation is present even in interaction modes without user prompting, such as in GitHub Copilot, which includes automated suggestions. In fact, such interaction modes may arguably make output evaluation more challenging due to the need to infer the intent behind systems' suggestions [74]. Novices

[10] The usability challenges of prompting make it a key target for explainability, as per §4.2.1.

[11] Confidence in the system's abilities is touched upon in §4.2.1 on explainability. Note also that self-confidence in output evaluation and confidence in the system's ability are both distinct from users' self-confidence in their *prompting* ability, discussed above in §3.1. Prompting and output evaluation influence each other as users iterate on their task.

[12] The two aspects of confidence can be independent. Having a well-calibrated confidence does not necessarily imply high confidence resolution. One could be well-calibrated on average (e.g., one's overall level of confidence matches one's overall level of accuracy) but still have poor resolution (i.e., one's confidence level does not vary much between correct and incorrect answers)[66].

[13] Related to output evaluation, confidence and metacognition have also been studied in the context of phishing detection [29].

in a domain or in GenAI may be particularly vulnerable, as one participant commented on long code suggestions: *"if you do not know what you're doing, it can confuse you more"* [142]. Given the importance of users' self-confidence for evaluating and relying on AI outputs, future research should measure and manipulate self-confidence during user-GenAI interactions across different interaction modes.

Output evaluation is relevant for many systems, such as search engines, but several aspects of GenAI pose unique challenges, which we discuss below: the *extensiveness* of GenAI's novel content output, the relative *ease* of novel content generation, GenAI's multiple, non-intuitive *failure modes*, and the challenge of obtaining *objective quality measures* for adjusting confidence in some workflows.

The extensiveness of GenAI's novel content output. Whereas prior research has focused on explicit AI advice or decisions, GenAI can produce (often extensive) content, such as entire emails, presentations, or software. Evaluating these outputs for quality therefore becomes far more important and effortful (in terms of cognitive load) compared to 'auto-complete' phrase suggestions or intelligent code completion [156, 157]. How this will affect users' self-confidence and AI reliance remains unclear, but metacognition research suggests that increased effort requirements may discourage users from appropriate evaluation [4]. Ackerman [2] found that people's internal confidence threshold for solving reasoning problems decreases as the required effort increases; that is, *"when problems took longer to solve, participants appeared to compromise on their confidence criterion, and were willing to provide solutions with less confidence"*. Worryingly, this persists even when participants are given the option to give up and respond "I don't know" [2]. Likewise, end-user programmers have been reported to "eyeball" the AI outputs of natural language queries, which some suggest may deepen the existing over-confidence that such users have in their programs' accuracy [156, 169].

Future research should examine how the effort of evaluating GenAI output (in terms of length or complexity) affects users' self-confidence in their output evaluation, the accuracy of their evaluation, and their ultimate reliance on GenAI.

The relative ease of novel content generation. The relative ease with which GenAI can produce extensive output may also affect output evaluation and reliance via potentially misleading cues that people implicitly rely on to update their confidence and guide their subsequent metacognitive control [4]. One relevant type of cue—'processing fluency', *"the subjective ease of with which a cognitive task is performed"* [3, 201]—can influence people's confidence in information accuracy. For example, answers to various problems are judged as more correct simply if they are displayed faster to participants after problem descriptions [185]. It can also affect people's confidence in their memory: the ease with which a memory is retrieved increases participants' confidence in their later remembering, even though, objectively, easier retrieval was associated with worse future memory performance [16].[14] Critically, this effect extends to technology use: faster online information search retrieval increases participants' confidence in their subsequent memory of that information, despite no apparent causal relation between the two aspects [172]. The mere *use* of technology, such as online information search, can also inflate people's confidence in their knowledge [53, 58, 63].

Analogously, the ability of GenAI systems to quickly and easily generate extensive content may serve as a cue that misleadingly increases users' confidence, not only in the output itself, but also in their own ability to evaluate it. More importantly, changes in confidence can affect people's approach to evaluating GenAI output. By increasing people's confidence, such cues can affect their metacognitive control, leading people to decrease the effort they invest into further deliberate processing, as measured by thinking time and changes-of-mind [4, 183, 184]. Confidence similarly influences reliance on external reminders and information-seeking [22, 49] (see also §3.3.2).

Future research should systematically investigate how aspects of GenAI output (e.g., the speed at which it's produced, or its verbal fluency, in the case of text) can serve as cues that influence users' confidence in the output and their ability to evaluate it, as well as the effort they ultimately invest into evaluation.

Multiple, non-intuitive failure modes of GenAI. Users' confidence and their ability to adjust it may also be challenged by the fact that GenAI tools can have multiple and often non-intuitive failure modes [33, 157]. For instance, they can introduce subtle, non-intuitive errors that a human would not introduce, further complicating evaluation [156]. As it stands, this requires developing an expertise and a well-adjusted confidence that is distinct from existing domain expertise, with, for example, *"developers [needing] to learn new craft practices for debugging"* [156]. Moreover, as noted, GenAI models are non-deterministic [136]. This is arguably a necessary trade-off within current GenAI systems, as it enables diversity of output [136, 156], yet it exacerbates the challenge of confidence adjustment, particularly when working iteratively across prompting and output evaluation (as per §3.1). How confident users should be in their evaluation ability, and how much effort they should invest in evaluation, partly depends on how much non-determinism they can expect in the output. Indeed, manufacturing designers co-creating with GenAI were, *"unable to determine whether... design features were intended or caused by algorithmic glitches"* [70]. More broadly, as per §3.1, output failures can be attributed to the user's prompt or parameter settings, or the system's non-determinism or training data, without an obvious way to disentangle these, further complicating confidence adjustment, particularly for non-expert users [157, 197].

Future work should examine how different reasons for output failures affect users' ability to appropriately adjust their confidence in their evaluation ability, and how that influences their evaluation of and reliance on GenAI output.

Obtaining appropriate measures for confidence adjustment. Adjusting confidence in output evaluation typically requires objective measures of performance for comparison (e.g., the number of errors a user correctly detected in the output), but the quality of generated content and its uses may be more difficult to evaluate objectively (e.g., consider how one would objectively evaluate the quality of an LLM-generated email) [87, 99]. The benefits of generated content may also be diffuse and indirect. For example, participants co-writing with an LLM found that seeing the LLM's suggestions was helpful even when they did not implement them

[14] The influences of processing fluency cues are examples of metacognitive experiences.

[207]. This implies subjectivity in the workflow which, although valid, makes it challenging to adjust one's confidence. Even with use-cases that involve subjectivity, such as creative tasks, users need to adopt an appropriate reliance strategy, which requires well-adjusted confidence. For example, given the ease of idea generation with GenAI, how can users be confident that the ideas generated by such systems are in fact helpful for their ideation process, rather than merely *feeling* like they are helpful?

Future work should develop more varied objective measures of output quality, and explore how user-provided subjective measures of quality can support users' ability to adjust their confidence in output evaluation (e.g., by considering the self-consistency of their reports [87]).

3.3 Automation strategy and generative AI workflows

Beyond the metacognitive demands implied in the local interaction with GenAI, the generality of GenAI—its applicability to a wide range of tasks[157]—poses a higher-level question to end-users about their workflow automation strategy: whether they should *"employ Generative AI, how, and how much is the utility of incorporating generated contents compared to conventional approaches"* [34]. Sarkar [154] describes this change as a shift from production to 'critical integration', where *"the output of AI systems will need to be integrated into a wider workflow involving human action"*, a process requiring critical evaluation of outputs. We argue that this imposes a distinct metacognitive demand on users that must make these decisions, akin to [149], who make a similar argument for digital storage and memory. That is, users must have self-awareness of the applicability and potential impact of using GenAI for their workflow; well-adjusted confidence in the ability to complete a task manually versus with GenAI; and metacognitive flexibility in adapting workflows to GenAI (Figure 2d-e). We first briefly summarize early evidence on how GenAI is impacting user workflows, and then discuss the role of metacognition in users' workflow automation strategy.

3.3.1 Early impact of generative AI on user workflows. Research on real-world GenAI workflows, primarily in AI-assisted coding, suggests that tools like GitHub Copilot alter users' workflows in diverse ways [166]. Although many changes may be positive and related to productivity boosts [40], we focus on the challenges to illustrate the demand for metacognition. In a sample of undergraduate students with programming experience, working with Copilot on realistic programming tasks was perceived to be challenging (although participants still strongly preferred it) [188]. Most relevantly, the generation of a long piece of code, particularly with errors, required participants to switch between coding, reading, and debugging, resulting in a high cognitive load (also reported in [14] and, in the domain of AI-assisted programming education, in [143]). Some users in [14] felt that Copilot was negatively restructuring their workflow by *"forcing them to jump in to write code before coming up with a high-level architectural design"*. They also reported writing more and differently worded comments for Copilot, which they then spent time deleting (unlike comments intended for humans).

Research in other domains points to similar potential challenges. Data scientists highlighted workflow integration as a key lever of control that determined the usefulness of AI assistance [119]. Writing workflows also substantially change with ChatGPT, with user time shifting from rough-drafting to editing [132], although the usability challenges that this brings remain to be explored.

Increased switching costs between automated and manual tasks, and automation-related restructuring of tasks in often unproductive ways have been studied in the human-automation interaction field for decades as the "ironies of automation" [12, 166]. Automated system design has adhered to best practices in human factors engineering to mitigate the impact of these challenges in specific contexts, such as driving. However, current GenAI systems present two key differences: they are applicable to a wide range of tasks ('generality' [157]), and as a result, they are also widely available to users with different levels of domain expertise, system training, and workflow standardization, in line with broader automation trends [82]. Thus, current systems shift the task of managing one's automation strategy to the user, who may lack expertise or training, leaving them to manage their attention and re-structured workflows as they see fit—a distinct metacognitive demand of GenAI. Broadly, these changes pertain to *understanding* and then *adapting* one's workflows. We discuss each of these below.

3.3.2 Understanding one's workflows: self-awareness and confidence adjustment. One key question that pertains to users' automation strategy is *whether* to automate a certain task. Inappropriate reliance on GenAI may result in lost productivity, increased risk of errors, or potential de-skilling [23]. Users must therefore have self-awareness of the applicability of GenAI for their workflow, and well-adjusted confidence in their ability to complete the task manually versus with GenAI [156] (Figure 2d). Put simply, it requires answering a version of the following questions: *do I know whether an available GenAI system can help my workflow; do I know how to work with it effectively in the context of my workflow; and how confident am I in this knowledge?* [114]. In end-user programming, this is known as the 'attention investment' problem, in which users must conduct a cost-benefit analysis to decide whether the potential attention costs saved from programming a manual task outweigh the attention costs of implementing the program [20].

Early research suggests that some users, particularly novices in GenAI and/or the task domain, lack sufficient self-awareness and well-adjusted confidence for working effectively with GenAI systems. For example, programming students in [143] repeatedly spent time editing Copilot suggestions before abandoning them and moving on, or tried to coerce Copilot to provide a correct suggestion, two unproductive interaction patterns that suggest potential over-reliance on GenAI. Similarly, some less experienced programmers in [14] were particularly excited about Copilot and would over-rely on it before manually attempting any of the tasks themselves. When compared with the relative absence of such interactions among experienced developers (e.g., [14, 188]), these interaction patterns illustrate a potential lack of metacognitive self-awareness and confidence in managing one's workflows. However, Kazemitabaar et al. [89] found no evidence of over-reliance among novice programmers when learning programming using GenAI.

The above reports are limited to short study contexts, focusing only on programming. Further in-depth research is needed on the impact of GenAI on realistic workflows, including understanding the role of user self-awareness and confidence, particularly for use-cases outside of programming.

Deciding to rely on GenAI is a form of 'cognitive offloading'—the use of tools external to the mind (e.g., calendars), to reduce the cognitive demand of a task (e.g., remembering an event) [148]. With GenAI, although the intent is often (but not always) to produce an external artefact, many cognitive processes that are traditionally involved in such production are at least partly 'offloaded' to GenAI, such as ideation, memory retrieval, and reasoning. For example, although users prompt systems with instructions to generate text, it's the systems which often generate ideas, retrieve relevant information, and structure it into arguments. Psychological research has explored how metacognition affects people's decisions to engage in cognitive offloading, and can therefore inform our understanding of the metacognitive demands pertinent to users' automation strategy [69, 148, 161]. Studies find that people's self-confidence in task performance or knowledge is a strong determinant of their use of external reminders [22, 68, 79], and search for external information [49, 158]. That is, a lower self-confidence in one's abilities is associated with more cognitive offloading. The above GenAI user studies demonstrate a similar pattern, where less experienced users were more likely to show patterns consistent with over-reliance on tools like Copilot. Critically, studies on cognitive offloading show that even when accounting for people's objective performance on a task, their *subjective* self-confidence still influences the decision to engage in cognitive offloading [22, 68].

Future work should explore how subjective self-confidence relates to users' automation strategies with GenAI, and how users, particularly novices, can be supported in having increased self-awareness and a well-adjusted confidence to ensure appropriate reliance on GenAI (see also §4).

3.3.3 Adapting one's workflows: metacognitive flexibility. Alongside self-awareness and confidence, working with GenAI requires metacognitive flexibility to be able to effectively adapt one's workflow (Figure 2e). For example, users should be able to recognize when and how the use of GenAI interferes with their workflow, resulting in a net productivity loss, and adjust accordingly. As discussed above, the challenges that some Copilot users faced suggest an under-development of this domain-specific metacognitive ability. Conversely, emerging evidence suggests that some experienced users do employ metacognitive flexibility in their workflows with GenAI. For example, some users with prior Copilot experience disable it entirely due to excessive disruption to their workflows [14]. In [103], 26 percent of surveyed programmers cite the distracting nature of GenAI suggestions, and 38 percent cite the time-consuming nature of debugging or modifying generated code, as 'very important' reasons for avoiding tools like Copilot. Certain data scientists in [119] similarly expressed skepticism towards AI assistance, particularly for difficult-to-understand generated code. Likewise, many manufacturing designers in [70] who struggled with GenAI ultimately avoided it altogether in their process.

Other users take a more nuanced approach: *"'I turned off auto-suggest and that made a huge difference. Now I'll use it when I know I'm doing something repetitive that it'll get easily, or if I'm not 100 percent sure what I want to do and I'm curious what it suggests. This way I get the help without having it interrupt my thoughts with its suggestions"'* [156]. More broadly, only about 20-30 percent of Copilot suggestions are accepted by users [211].

This evidence also hints at relevant differences between system interfaces and interaction modes that should be considered in the context of metacognitive demands and adapting workflows. Automated suggestions reflect tighter integration between manual work and GenAI, and no metacognitive demands associated with prompting, yet they nevertheless present challenges such as interruptions. In contrast, prompt-based interactions enable more user control over workflows but present metacognitive demands (as per §3.1.1). User-controlled suggestions may be a middle-ground, but present their own challenges in terms of inferring system intent (as per §3.2.1). Deciding between these approaches as a user may contribute to the metacognitive demand associated with determining automation strategy (see also [171]).

This is not to suggest that users' approaches to workflow adaptation above are necessarily optimal for productivity, but rather that they reflect self-awareness and confidence in experienced users about how GenAI impacts their workflow, and the exertion of metacognitive flexibility in an effort to change this. As Sarkar et al. [156] conclude, these emerging ad hoc strategies hint at *"a new cognitive burden of constantly evaluating whether the current situation would benefit from LLM assistance"*—a burden that we identify as distinctly metacognitive. Future work should characterize the role of metacognitive flexibility in adapting one's workflows across different GenAI interfaces and interaction modes.

4 ADDRESSING THE METACOGNITIVE DEMANDS OF GENERATIVE AI

The metacognitive demands posed by GenAI can be addressed in two complementary ways: (1) *improving users' metacognition* via metacognitive support strategies that can be integrated into GenAI systems, and (2) *reducing the metacognitive demand* of GenAI systems by designing task-appropriate approaches to explainability and customizability. The distinction between the two approaches is not clean-cut, yet helps frame the design space.

Multiple lines of evidence suggest that metacognition can be improved, and that individuals who are supported in specific metacognitive monitoring or control abilities can significantly improve their performance in metacognitively demanding tasks [45, 52]. This applies across different age groups (from children to adults) [8, 39, 48], tasks (e.g., lecture comprehension or mathematical reasoning) [27, 80, 91, 96], time-scales (i.e., immediately as well as delayed) [121], and learning settings (i.e., solitary as well as social) [137]. As such, interventions to improve the metacognition of users working with GenAI could be one effective way of meeting the demands of these systems. This includes embedding metacognitive support strategies—for example, supporting users' planning and self-evaluation—directly into GenAI systems. Interventions can be adapted to the metacognitive abilities and GenAI experience of each user to provide the appropriate level of support, making productive use of GenAI's model flexibility and generality.

Table 2: Open research questions for understanding the metacognitive demands of GenAI

Area	Research questions	Example measures of metacognition
Prompt formulation	How does self-awareness and task decomposition ability moderate users' ability to control systems across interaction modes (e.g., diegetic vs. non-diegetic prompting), task contexts (e.g., creating a novel output vs. editing an existing artifact), and domains (e.g., writing vs. programming)?	Think-aloud, self-report protocols, SRL microanalysis
Prompt iteration	How do different aspects of GenAI systems (e.g., non-determinism, model flexibility) impact users' ability to adjust their confidence in their prompting ability and flexibly adapt their prompting strategy and mental model of GenAI systems?	Prospective self-ratings of confidence, SRL microanalysis
Output evaluation	How does users' self-confidence in a task domain or with GenAI influence their output evaluation and reliance across different interaction modes?	Prospective and retrospective self-ratings of confidence
	How does the cognitive load associated with evaluating GenAI outputs (e.g., in terms of output length or complexity) affect users' self-confidence, the accuracy of their evaluation, and their ultimate reliance on AI?	Retrospective self-ratings of confidence, meta-*d*' estimates
	How do aspects of GenAI output (e.g., the speed at which it is produced, its verbal fluency in the case of text) serve as heuristic cues that influence users' confidence in the output and their evaluation ability, as well as the amount of effort they invest into evaluation?	Retrospective self-ratings of confidence, meta-*d*' estimates
	How do different reasons for output failures affect users' ability to adjust their confidence in their evaluation ability, and how does that influence their strategies for evaluating and relying on GenAI output?	Prospective judgments of learning, retrospective reflective journals
	What are useful objective measures of quality for long and/or multidimensional outputs, and how can user-provided subjective measures of quality support users' ability to adjust their confidence in output evaluation?	Self-ratings of confidence
Understanding workflows	How does subjective user confidence and self-awareness in a domain and/or in their ability to work with GenAI relate to users' automation strategies with GenAI?	Retrospective self-ratings of confidence, SRL microanalysis
Adapting workflows	What is the role of metacognitive flexibility in adapting one's workflows across different GenAI interfaces and interaction modes?	SRL microanalysis

On the other hand, GenAI systems can be designed to reduce their metacognitive demand. One area ripe for this approach is explainability. Designing human-centered explainable AI (HCXAI) has been an important focus in human-AI interaction research [57, 104, 175], but the model flexibility, generality, and originality of GenAI systems poses further challenges, as per §3. Yet these same features of GenAI provide an opportunity to support HCXAI, particularly when considering it through the lens of metacognition. Alongside explainability, the customizability of GenAI systems is another lever to reduce metacognitive demand. Current GenAI systems provide many parameters to users, both explicitly (as settings), and implicitly (as prompting strategies). Finding appropriate ways to surface these can reduce metacognitive demand.

Below, we discuss how to improve users' metacognition using three types of metacognitive support strategies that can be employed in GenAI systems: planning, self-evaluation, and self-management. After discussing the range of possible strategies for each kind of metacognitive support, we provide a figure showing a hypothetical example of how they might be used in a scenario within an existing GenAI system (analogously to [26]). We then turn to how systems can be designed to reduce metacognitive demand, focusing on explainability and customizability. We provide examples from research on metacognition interventions and existing prototype studies and suggest opportunities for further research. Lastly, we briefly discuss the importance of managing the cognitive load associated with metacognitive interventions.

4.1 Improving user metacognition

4.1.1 Planning. Planning is a task-oriented metacognitive strategy entailing both the definition of clear goals (i.e., self-awareness) and devising a comprehensive approach for achieving them by breaking them down into smaller, manageable steps (i.e., task decomposition) [21]. Planning-related interventions can support both of these aspects as users work with GenAI systems.

As discussed below in §4.1.2, self-evaluation interventions can help users reflect on their task goals and approaches to task decomposition [74]. However, tasks may nevertheless be complex or ambiguous, often requiring gathering, organizing, and synthesizing information in a nonlinear manner distinct from the linear conversational interfaces in most GenAI systems today. For example, people can engage in multi-level planning, where hours, days, and weeks have to be considered simultaneously [6]. More flexible interfaces can support the crafting of a complex prompting strategy by enabling open-ended exploration of task goals and the relationships between them during the planning process. *Sensecape* is such an interface for LLMs that uses multilevel abstraction and visuo-spatial organization to support exploration and sensemaking during LLM interactions [173]. Similar systems have been shown to improve users' planning and other metacognitive processes [38]. When applied to the prompting and output evaluation process itself, these approaches can make users aware of where they are in a task. They can also help them encode information in personalized representational schemas [141], which can help users understand the underlying mechanisms of a GenAI system, its capabilities, and failure points. Such an understanding can in turn mitigate the risks of inappropriate evaluation of, and confidence in, AI-generated output.

Planning-related interventions can also support users directly in task decomposition, improving prompt effectiveness via more explicit and discrete instructions. One promising approach is 'prompt chaining', which involves *"decomposing an overarching task into a series of highly targeted sub-tasks, mapping each to a distinct LLM step, and using the output from one step as an input to the next"* [203, 204]. Alongside improving the LLM's ability to execute complex tasks, chaining helped participants *"think through the task better"*, and thereby make more targeted edits to improve their prompting [204]. Chaining also increased users' self-awareness of their goals: the ability to decompose tasks led some participants to create more generalizable outputs better suited to their broader goals [203].

Planning can also help users address the 'fuzzy abstraction matching' problem [156], that is, translate their goals and intentions into executable actions—a form of externalization where 'tacit' knowledge is made into explicit prompts [130]. This can be supported through feedforward design [34] (as distinct from feedback [191]): inviting an action and communicating what exactly the user can expect as a result.[15] For example, feedforward can be used to inform users that a vague, high-level prompt is unlikely to achieve their task before they submit it. As Vermeulen et al. [191] argue, *"the more complex a system or interaction context gets, the larger the need will be for elaborate feedforward in order to aid users in achieving their goals"*. Prompt chaining is an example of a sequential feedforward approach, surfacing inputs, outputs, and the underlying prompt structure for users to explore [204]. However, feedforward information can also increase cognitive load [42], so the right balance, particularly for complex systems, remains to be explored. The *Prompt Middleware* framework uses feedforward at different levels of complexity to guide users towards effective prompts and scaffold domain expertise into the process [113].

Figure 3 provides a hypothetical example of a planning-focused metacognitive intervention that could be implemented in a conversational interface such as ChatGPT. Rather than requiring an entire chain structure with full control over inputs and outputs, the benefits of chaining could be derived by surfacing a relevant set of key questions to users in a more accessible format.

4.1.2 *Self-evaluation.*

Self-evaluation involves enabling users to reflect on their knowledge, strategies, and performance, and their respective level of confidence in these. Interventions that cue users to reflect on their goals and strategies via question prompts or conversational interfaces have been shown to improve outcomes in the workplace [92, 122, 209], education [51], and other domains [17, 147]. GenAI systems, with their model flexibility and generality, have the potential to adaptively nudge this kind of self-evaluation at key moments during user workflows, effectively acting as a coach or guide for users [78].

Self-evaluation can be used to support effective prompting. For example, Gmeiner et al. [70] employed human experts to guide designers during their interaction with GenAI systems; users appreciated critical questions that guided self-evaluation and thought this improved prompting. Gmeiner et al. suggest that GenAI systems can proactively offer similar context-aware self-reflection prompts to support users in thinking through problems. Self-evaluation interventions can rely on a range of design elements to support users in clarifying their task goals, including temporal aspects (e.g., considering past tasks or broader project timelines), comparison (e.g., with similar tasks), and discovery (e.g., through re-framing tasks) [17].

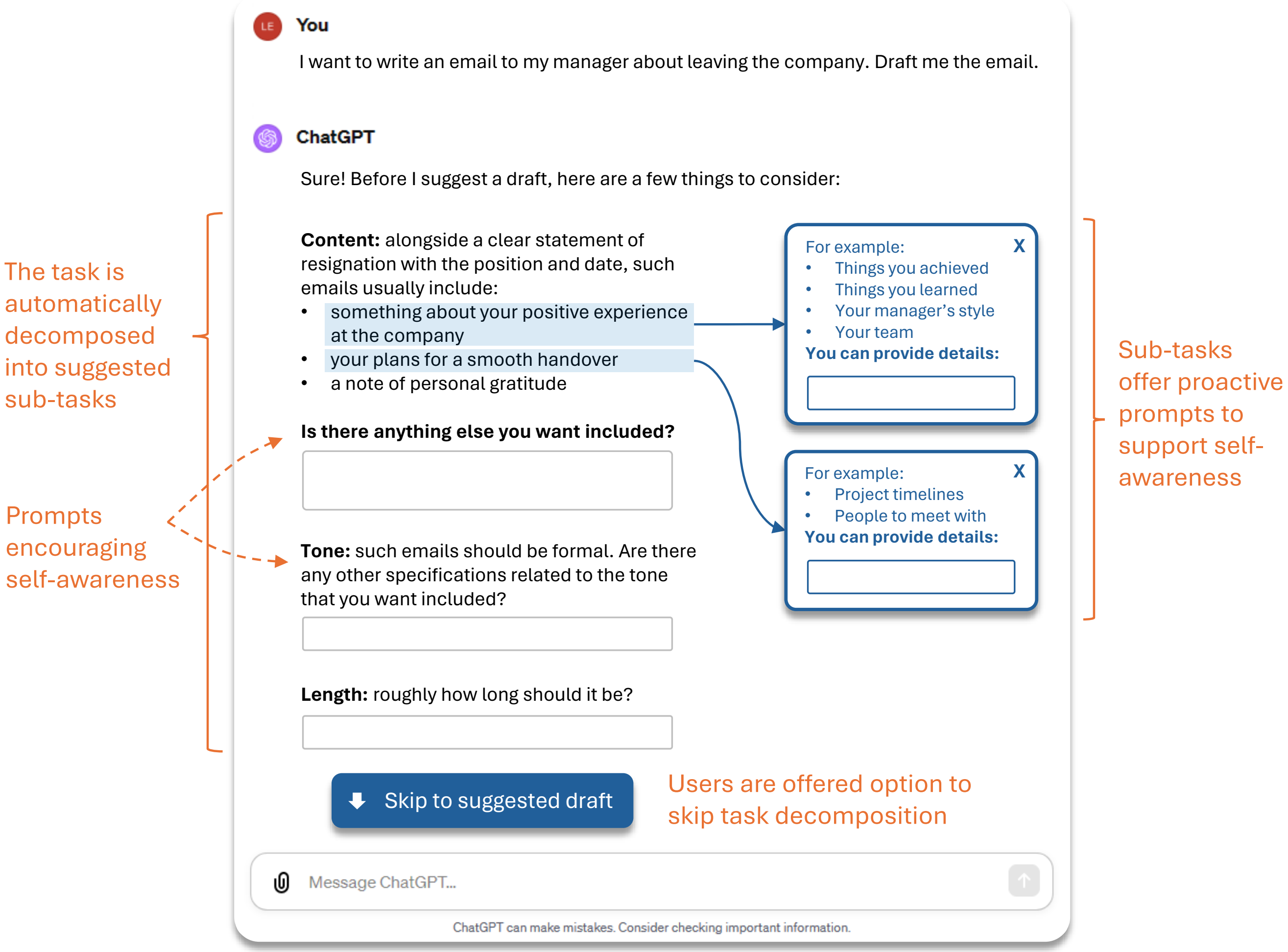
Self-evaluation during the output stage can include interventions that surface previous outputs and ask users to 'think aloud' about their thought processes, promoting self-awareness in 'real time' and supporting users in detecting hallucinations in GenAI outputs [84]. For example, simple textual prompts promoting self-awareness significantly decreased participants' susceptibility to incorrect (albeit realistic) information [153]. Likewise, self-explanation prompts improved students' accuracy in evaluating their own understanding of information [201]. GenAI systems could also probe for and proactively respond to user uncertainty. Even including a 'not sure' option for users auditing LLMs, can enable them *"to reflect on the task specification and the appropriateness of the tests considered"* [146].

Self-evaluation is particularly promising for augmenting GenAI explainability, where it can help increase users' receptiveness to explanations that seek to update their mental models of GenAI systems [4, 171] (see §4.2.1 for more on explainability). For example, self-reflection probes can ask users to reflect on their mental model and associated confidence [97].

GenAI systems also provide opportunity for *interactive* metacognitive support, encouraging self-evaluation and suggesting adjustments to metacognitive strategies based on context [202]. For instance, systems could interactively guide users through the steps of a problem rather than simply providing a solution [70]. Indeed, the success of pair programming often stems from the extensive verbalization between programmers, and resultant self-evaluation, rather than the division of labor [75]. Interactions with GenAI systems could replicate this with a similar level of reflective depth. Enabling users to critically evaluate AI outputs can build self-awareness by forcing users to rationalize their decisions to the GenAI system [156], enhance metacognitive flexibility by providing users with different perspectives on their task [15], and adjust confidence by adapting explanations to the users' needs and the models' confidence [110]. Moreover, it can also provide critical feedback to GenAI systems themselves.

However, there are also potential pitfalls in interactivity. As per §3.2, processing fluency can lead to inflated confidence without necessarily improving objective accuracy. The design of intuitive interfaces for GenAI systems might therefore inadvertently give users a misleading sense of competence, increasing the risk of errors (likewise discussed in cognitive psychology [153]). Designers

[15] This is also a form of explanation; see §4.2.1.



You

I want to write an email to my manager about leaving the company. Draft me the email.

ChatGPT

Sure! Before I suggest a draft, here are a few things to consider:

Content: alongside a clear statement of resignation with the position and date, such emails usually include:

- something about your positive experience at the company
- your plans for a smooth handover
- a note of personal gratitude

For example:
- Things you achieved
- Things you learned
- Your manager's style
- Your team

You can provide details:

For example:
- Project timelines
- People to meet with

You can provide details:

Is there anything else you want included?

Tone: such emails should be formal. Are there any other specifications related to the tone that you want included?

Length: roughly how long should it be?

The task is automatically decomposed into suggested sub-tasks

Prompts encouraging self-awareness

Sub-tasks offer proactive prompts to support self-awareness

Users are offered option to skip task decomposition

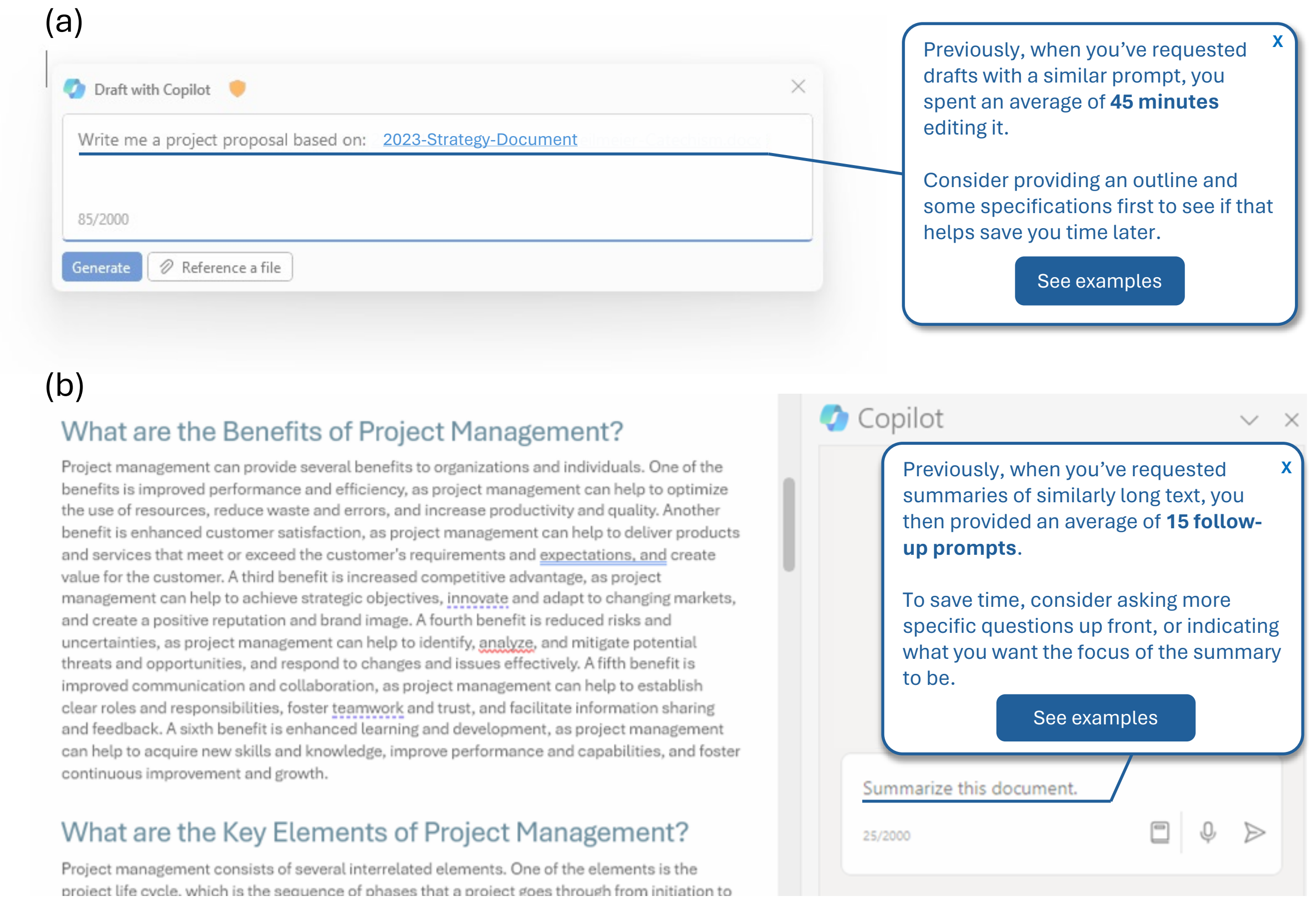
ChatGPT can make mistakes. Consider checking important information.

Figure 3: Hypothetical example of a planning-focused metacognitive intervention built into ChatGPT. After the user specifies a task, the system automatically comes up with a decomposed, step-by-step guide for completion (left side of the figure). This could be aided by further proactive prompting, giving concrete examples of how sub-tasks could be solved (right side of the figure). An option to skip the decomposition step (bottom of the figure) minimises unnecessary cognitive load if decomposition is not required.

therefore need to carefully consider how to improve processing fluency without leading to overconfidence, for example, by including periodic checks that challenge users' assumptions or solutions [94]. This speaks to 'seamful' design, which leads users to pause or reflect on their engagement with technology by emphasizing *"configurability, user appropriation, and revelation of complexity, ambiguity or inconsistency"* [81, 155, 196].

Figure 4 provides a hypothetical example of a metacognitive intervention focused on self-evaluation. To support effective prompting, the user's prior history is leveraged to provide personalized suggestions to improve a generic prompt in this case, and potentially teach the user to include more detail in future prompting.

4.1.3 Self-management. Self-management involves the strategic management of variables like time, setting, and workflow, and is therefore an important focus for metacognitive support strategies for GenAI users. These considerations are not just arbitrary choices but can be informed by a blend of user telemetry trends and explicit user requests. By developing systems that are context-aware, users can be served AI-generated content or prompts at opportune moments during their workflows [74]. For example, coding assistance systems can detect when a user is in a state of flow ('acceleration') or problem-solving ('exploration'), adapting code suggestions accordingly and providing feedback to the user [14]. Likewise, during highly sensitive tasks or crucial time periods, the system could trigger heightened user engagement or display more salient reminders to critically evaluate the AI-generated output, promoting



Write me a project proposal based on: 2023-Strategy-Document

Previously, when you’ve requested drafts with a similar prompt, you spent an average of **45 minutes** editing it.

Consider providing an outline and some specifications first to see if that helps save you time later.

What are the Benefits of Project Management?

Project management can provide several benefits to organizations and individuals. One of the benefits is improved performance and efficiency, as project management can help to optimize the use of resources, reduce waste and errors, and increase productivity and quality. Another benefit is enhanced customer satisfaction, as project management can help to deliver products and services that meet or exceed the customer’s requirements and expectations, and create value for the customer. A third benefit is increased competitive advantage, as project management can help to achieve strategic objectives, innovate and adapt to changing markets, and create a positive reputation and brand image. A fourth benefit is reduced risks and uncertainties, as project management can help to identify, analyze, and mitigate potential threats and opportunities, and respond to changes and issues effectively. A fifth benefit is improved communication and collaboration, as project management can help to establish clear roles and responsibilities, foster teamwork and trust, and facilitate information sharing and feedback. A sixth benefit is enhanced learning and development, as project management can help to acquire new skills and knowledge, improve performance and capabilities, and foster continuous improvement and growth.

What are the Key Elements of Project Management?

Project management consists of several interrelated elements. One of the elements is the project life cycle, which is the sequence of phases that a project goes through from initiation to

Previously, when you’ve requested summaries of similarly long text, you then provided an average of **15 follow-up prompts**.

To save time, consider asking more specific questions up front, or indicating what you want the focus of the summary to be.

Summarize this document.

Figure 4: Hypothetical example of a metacognitive intervention focused on self-evaluation built into Microsoft Copilot. In (A), the user provides a highly unspecified prompt to the system for writing a proposal. Based on a neutral assessment of similar prompting history, the GenAI system suggests reducing editing time by reflecting on more strategies. In (B) the user provides a highly unspecified prompt for summarizing. Based on a neutral assessment of previous interactions, the GenAI system suggests to limit interactions by suggesting to reflect on the user’s more specific goals and intentions for this summary. These appear as suggestions next to the main chat window and can be closed if not wanted.

self-awareness and adjustment of confidence in output. Another approach is designing the complexity of AI-generated content according to the cognitive load experienced by the user [177]. This could involve dynamic adaptations, such as recognizing implicit intent [34], providing summaries when the user is overwhelmed, or escalating the complexity when the user demonstrates high proficiency and engagement. Supporting self-management efficiently also depends on task demands [74]. For example, [1] and [164] suggest that in the context of solving logical puzzles, intelligent tutors should offer backwards-oriented workflows (e.g., using prompts to encourage thinking about the negation of the actual solution), rather than focusing on forward-oriented workflows. In a data science context, Gu et al. [74] propose that GenAI tools can offer *“a ‘think’ mode for specific planning suggestions, a ‘reflection’ mode for connecting decisions and highlighting potential missed steps, and an ‘exploration’ mode for higher-level planning suggestions”.*

Deciding when to present *any* GenAI support is another design choice affecting self-management. In AI-assisted decision-making, Steyvers and Kumar [171] distinguish between AI support provision that is on-demand (user-requested) and sequential (occurring after a user makes an independent decision), among others. Apart from facilitating engagement at opportune moments, the sequential paradigm is presumed to encourage independent reflection by the user. Park et al. [138] similarly argue that “slower” interfaces especially enable these benefits, as the waiting time often gets used for reflective thinking about the task at hand (see also [147]). Alternatively, workflows can be more dynamic. For example, pathology requires highly specialized, moment-to-moment judgments; in this context, the user capability to control and modify search algorithms on-the-fly can be particularly beneficial [28].

Figure 5 provides a hypothetical example of a metacognitive intervention focused on self-management and self-evaluation. During coding, a system might encourage the user to reflect on whether all relevant parameters are included, check on whether complex code is understood (especially useful if code has been imported from other sources and may impact critical aspects of operation), or the broader work context. Critically, it offers options to ignore the suggestion, schedule it for later, or change proficiency settings (i.e., self-confidence).

4.2 Reducing metacognitive demands

In addition to improving users' metacognition during their interaction with GenAI, systems should be designed to *reduce their metacognitive demands*. Target areas for this include the explainability and customizability of GenAI systems.

4.2.1 Explainability. We adhere to the definition of explainability as that which enables *"people's understanding of the AI to achieve their goals"* [174]. To this point, §3 demonstrated how users struggled to understand GenAI systems and achieve their goals due to GenAI's metacognitive demands. We focus on those using GenAI systems as tools in their workflow (in contrast to, e.g., those engaging solely with GenAI system outputs, or overseeing regulatory aspects), in line with the context-specificity of explainability advocated by research on HCXAI [57, 104, 175].

From the perspective of metacognitive demand, by providing contextual and performance information alongside the system inputs and outputs, explainability should help partly *offload* metacognitive processing from users' minds and onto the system interface. As we illustrate below, explainability approaches can surface the information necessary for adjusting confidence in prompting, output evaluation, and automation strategy. Moreover, by providing actionable information, explainability can enable users' self-awareness and metacognitive flexibility [116].[16]

Explainability for GenAI systems should help users adjust their confidence in their ability to prompt, evaluate outputs, and determine their automation strategy, particularly given the multiple, non-intuitive failure modes common to GenAI systems, and other challenges [104]. For example, explanations that map each aspect of an output to aspects of the prompt (e.g., using attention visualization [179]), and compare this to examples of effective prompts [25, 85], can help users disentangle issues with their prompt from those stemming from model performance, thereby supporting confidence adjustment for prompting ability. Likewise, 'co-auditing' a GenAI system by revealing the model's step-by-step actions (e.g., in spreadsheet software [105]) can enable users to understand exactly what's involved in a longer workflow [71], and help them adjust their confidence in their evaluation ability. For example, an explanation that introduces domain concepts or terminology unfamiliar to users can signal insufficient domain expertise to evaluate this output [165], and prompt further explanation. Finally, indicating model uncertainty [19, 174, 198], for example, by means of color-coding [167], and an explanation for that uncertainty [5, 190], can also help users disentangle the role of their prompting and output evaluation ability from that of models' capabilities, further supporting confidence adjustment [197].

The broader user workflow that encompasses the local interaction with GenAI constitutes an important usage context for explainability (see also [99, 174]). To this end, global explanations about model capabilities for a given task can help users adjust their confidence in their ability to complete the task manually versus with GenAI support (i.e., determining their automation strategy) [85]. Indeed, for AI-assisted coding, users requested information about overall output quality and runtime performance [174].

Explainability can also reduce the metacognitive demand for user self-awareness and metacognitive flexibility. For example, explanations about effective prompting strategies for a given task—most frequently requested by users in a GenAI-assisted coding system [174]—can help users translate their goals into actionable prompts or flexibly adjust their prompts (which can equally be viewed as supporting their metacognitive abilities, as in feedforward design [191]). Moreover, granular model uncertainty estimates, such as line-level highlighting of generated code, can support users in prioritizing their output evaluation [190, 198, 199], thereby enabling metacognitive flexibility. In sum, these approaches to explainability all aim to reduce the metacognitive demand of GenAI systems, enabling users to take concrete actions, including 'mental state' actions (e.g., confidence adjustment), system interactions (e.g., updating prompts), or actions external to the system (e.g., completing a task without GenAI) [116].

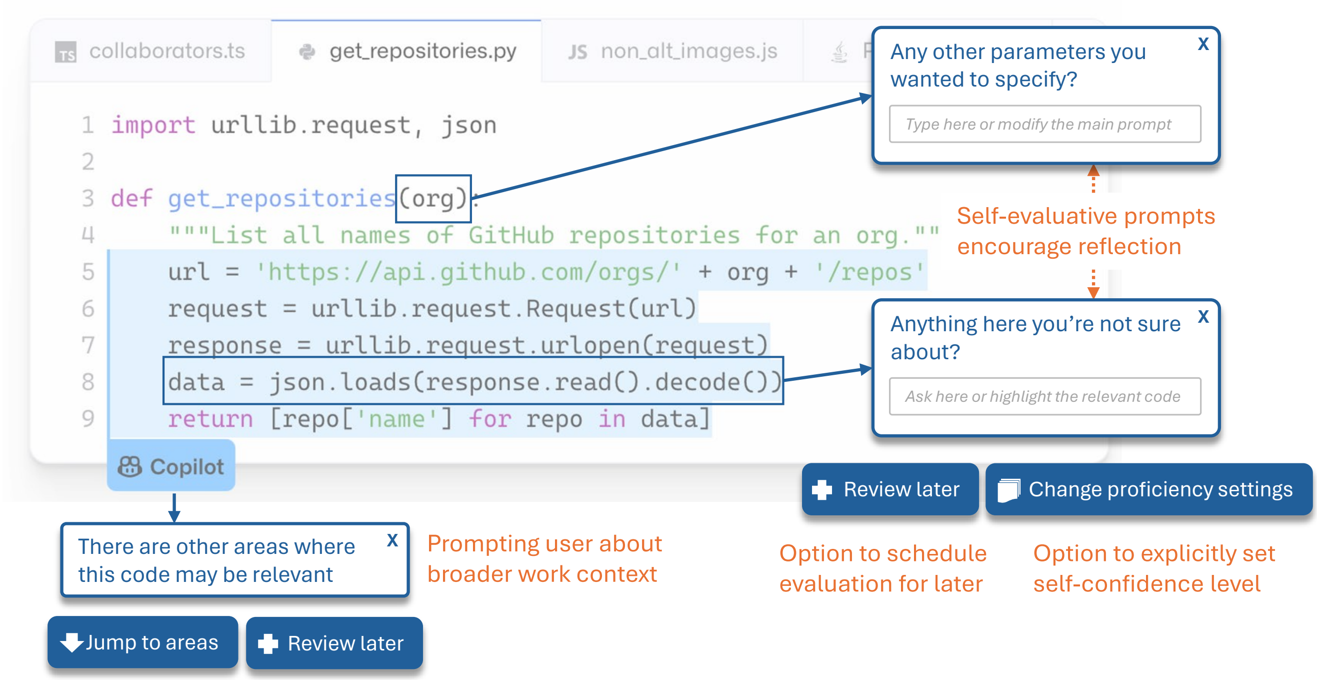
As suggested in §4.1.2, the above explainability approaches can be augmented via metacognitive self-evaluation interventions that encourage self-awareness. GenAI offers a unique opportunity to further augment these interventions through interactivity, as advocated in recent work [10, 100, 101, 123, 192], including for GenAI specifically [174]. Interactivity could be especially important for GenAI explainability, as due to GenAI's model flexibility, generality, and originality, users' explanation needs will be as diverse as their use-cases, outputs, and metacognitive abilities.

4.2.2 System customizability. The question of how much control to give users—system customizability, or how many 'knobs' a user can adjust—is another design choice that can moderate the metacognitive demands of GenAI [42, 119].

On one hand, increasing customizability can increase the demand for self-awareness (e.g., *"are any of the settings relevant to my task?"*), well-adjusted confidence (e.g., *"did any of the settings affect the output quality, or is it related to my prompting?"*), task decomposition (e.g., *"what is the right order to adjust settings for each of my subtasks?"*), and metacognitive flexibility (e.g., *"which settings should I adjust to improve my output, if any?"*). This may increase cognitive load, particularly for novice users. Indeed, in Perry et al. [139], half of the users did not adjust any model parameters, even though many produced insecure code using an AI assistant. To this end, the *Prompt Middleware* framework aims to reduce the demand to craft prompts from scratch, enabling users to choose limited customizability [113].

On the other hand, increased customizability can support metacognition, particularly for more advanced users. For example, consider the temperature setting, which determines the extent of non-determinism in GenAI outputs, and therefore the likelihood of hallucinations. Allowing users to change this setting can support more flexible and self-aware problem-solving in experienced users,

[16] In this sense, explainability can also be viewed as a metacognitive support strategy as per §4.1.


collaborators.ts
get_repositories.py
non_alt_images.js
X
Type here or modify the main prompt
X
Ask here or highlight the relevant code
Copilot
X
Review later
Change proficiency settings
Jump to areas
Review later

```python
import urllib.request, json

def get_repositories(org):
    """List all names of GitHub repositories for an org.""
    url = 'https://api.github.com/orgs/' + org + '/repos'
    request = urllib.request.Request(url)
    response = urllib.request.urlopen(request)
    data = json.loads(response.read().decode())
    return [repo['name'] for repo in data]
```

Any other parameters you wanted to specify?

Self-evaluative prompts encourage reflection

Anything here you're not sure about?

There are other areas where this code may be relevant

Prompting user about broader work context

Option to schedule evaluation for later

Option to explicitly set self-confidence level

Figure 5: Hypothetical example illustrating a metacognitive intervention focused on self-management and self-evaluation for coding in GitHub Copilot. During programming, the system can provide self-evaluation prompts to encourage user reflection on bugs and purpose of the code (right side of the figure). To decrease cognitive load if evaluation is not wanted, the user has the option to schedule evaluation for later, or set their own confidence level to increase or decrease the amount of suggestions (the 'proficiency settings' in the bottom right of the figure). The system could also prompt the user to think about the broader work context, for example whether this code-snippet may be relevant somewhere else in the overall code as well. The user has the options to ignore this suggestion, look at the other relevant areas now, or review them later (bottom left of the figure). To minimize cognitive load, the timing and frequency of prompts should be adapted to the users' preferences, expertise, and workflow.

by presenting them with different and perhaps surprising perspectives [139]. Allowing users to find a task-appropriate temperature setting that keeps the right balance between diversity and factuality of output, or constraining factuality in different ways (i.e., through automated post-processing and deterministic fact-checking) could therefore enable metacognitive flexibility and self-awareness. Customization may therefore increase users' trust in and satisfaction with output [135].

Other settings include the size of the shortlist from which output is sampled, as well as the size of the output itself. Increasing the size of the output window may demand better-adjusted confidence to evaluate and integrate more information. However, it can also support metacognition—by adjusting these parameters, users can work on understanding which part of the output should be used and how, potentially also increasing explainability. Initial self-reports such as [152] suggest that in order to find the optimal system settings for a task, users enter an interactive feedback loop with models, in which they clearly have to formulate their goals (promoting self-awareness), adjust their confidence in output evaluation, and flexibly adapt their workflow.

Where the balance lies between too much and not enough customizability needs exploring [34, 197]. Combining increased customizability with metacognitive support strategies (e.g., planning, self-evaluation, self-management) is a promising direction for further research.

4.3 Managing cognitive load while addressing metacognitive demands

There is a risk that strategies to address metacognitive demands may also increase cognitive load, due to the additional information that users have to process, such as self-reflective prompts, a set of sub-goals resulting from task decomposition, or model explanations [46]. The relationship between metacognition and cognitive load is an active research area [46, 163, 193, 194], but it is plausible that, although cognitive load may increase due to the *processing requirements* of many metacognitive support interventions, the improvement in metacognition may be accompanied by a simultaneous and larger reduction in cognitive load, resulting in a *net decrease* in cognitive load. Some studies show that metacognitive support does not increase overall cognitive load [109, 210] (see also

Table 3: Open research questions for addressing the metacognitive demands of GenAI

Area	Research questions	Suggested approaches
Supporting users' metacognition	How can GenAI systems increase users' self-awareness and task decomposition during prompting?	Explore self-reflection prompts, task decomposition support, open-ended exploration, feedforward design, and other planning interventions.
	How can GenAI systems incorporate self-evaluation interventions to support users in increasing their self-awareness and adjusting their confidence? Does this affect their automation strategy?	Explore proactive probing of users' uncertainty, prompting users to self-explain and reflect interactively, and outputting systems' confidence.
	How can GenAI systems incorporate self-management interventions to support users in determining their automation strategy and improving their self-regulation and metacognitive flexibility?	Explore automated task decomposition, detection of users' states, dynamic adaptation of output complexity, and prompting towards more structured and interactive usage of GenAI.
Reducing metacognitive demand	How can explainability reduce the metacognitive demand of GenAI, and what is the impact of interactive explanations?	Explore impact of surfacing interactive output explanations at different levels of complexity.
	How can understanding users' metacognitive abilities when working with GenAI systems advance approaches to explainability and updating of mental models?	Monitor metacognitive abilities during GenAI interactions and explore whether metacognitive interventions improve mental model updating.
	What is the optimal balance for GenAI system customizability to reduce the metacognitive demand, and how can it be combined with metacognitive support strategies?	Explore different levels of customizability across tasks and user proficiency levels, and their impact on task performance and metacognition.
Managing cognitive load while addressing metacognitive demands	How do metacognitive interventions affect cognitive load as users learn to interact with GenAI systems over time, and how should interventions optimally adapt or fade out?	Explore reducing or otherwise adapting interventions at different timescales as metacognitive proficiency and task performance increases.
	What are other ways to optimize the balance between addressing metacognitive demands and overall cognitive load?	Explore context-appropriate gamification of metacognitive interventions.

[194]). Most importantly, we hypothesize that improved metacognition should result in a net improvement in output quality.

Likewise, we propose that explainability, by partly offloading metacognitive processing from users and onto the system, should reduce the cognitive load *associated with metacognitive monitoring and control*. However, it may increase the cognitive load *associated with processing explanations* [46]. To the latter point, some interactive explanations have been found to increase cognitive load [18]. Ultimately, however, as users adapt to working with explainable GenAI systems, we hypothesize that the result should be a net reduction in cognitive load [134]. As noted above, system customizability involves a similar tension between cognitive load and metacognition.

Training effects over time may be key, as users may gradually internalize the metacognitive strategies, explanations, or customization settings, and no longer need to rely on external prompts [46]. Accordingly, metacognition-related cognitive load should decrease, although it is less clear whether cognitive load associated with the *processing* of external prompts also sufficiently decreases. To this point, [133] found that adaptive and gradual fading out of metacognitive prompts produced the largest performance benefits, as it provided time for students to internalize metacognitive strategies, while ultimately reducing the cognitive load associated with processing now-irrelevant external prompts (see also [134]). The same might be true for some types of explanations as well (e.g., global explanations). Future research should study how metacognitive interventions affect cognitive load as users learn to interact with GenAI systems over time, and how to optimally adapt or fade out interventions over time. It is also important to explore other ways to optimize the balance between addressing metacognitive demand and overall cognitive load (e.g., through gamification [165]).

Lastly, and perhaps somewhat controversially, we highlight the value of 'seamfulness' in interface design for helping users reflect on their technology use (as per §4.1.2 and [81, 196]). This idea can be extended to question the 'doctrine of simplicity', which assumes that interfaces should always be 'easy' or 'natural' to use [155]. We propose that *some* potential effort introduced by metacognitive support strategies and explanations may be justified, so long as these are well-designed and act ultimately in the service of improved metacognition and productivity with GenAI, a technology which promises to transform personal and professional work [155].

5 CONCLUSION

Russell [50] proposed that being literate in the "Age of Google" required a kind of 'meta-literacy'—knowing how to read the search interface, how to use it effectively, and what is even possible to search for. Analogously, as we offload more of our cognition to today's GenAI systems, the demand for our metacognition increases [149]. Designing truly human-centered GenAI systems [34, 104] means grappling with these metacognitive demands. Fortunately, a rich body of metacognition and cutting-edge HCI research can kickstart this effort. Equally, interaction with GenAI offers a powerful paradigm for advancing our foundational understanding of metacognition, paving the way for fruitful inter-disciplinary research. Finally, we reiterate that the perspective of metacognition, when considered with the unique features of GenAI—model flexibility, generality, and originality—presents an opportunity to realize what Alan Kay proposed as a *"grand collaboration"* with *"agents: computer processes that act as guide, as coach, and as amanuensis"* [88].

ACKNOWLEDGMENTS

We thank Siân Lindley, Andy Gordon, Sam Gilbert, and the anonymous reviewers for their constructive comments.

REFERENCES

[1] Mark Abdelshiheed, John Wesley Hostetter, Preya Shabrina, Tiffany Barnes, and Min Chi. 2023. The Power of Nudging: Exploring Three Interventions for Metacognitive Skills Instruction across Intelligent Tutoring Systems. https://doi.org/10.48550/arXiv.2303.11965 arXiv:2303.11965 [cs].

[2] Rakefet Ackerman. 2014. The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General* 143, 3 (2014), 1349–1368. https://doi.org/10.1037/a0035098 Place: US Publisher: American Psychological Association.

[3] Rakefet Ackerman. 2019. Heuristic Cues for Meta-Reasoning Judgments: Review and Methodology. *Psihologijske teme* 28, 1 (May 2019), 1–20. https://doi.org/10.31820/pt.28.1.1 Publisher: Filozofski fakultet u Rijeci.

[4] Rakefet Ackerman and Valerie Thompson. 2017. Meta-Reasoning: Shedding meta-cognitive light on reasoning research. 1–15.

[5] Mayank Agarwal, Kartik Talamadupula, Stephanie Houde, Fernando Martinez, Michael Muller, John Richards, Steven Ross, and Justin D. Weisz. 2021. Quality Estimation & Interpretability for Code Translation. https://doi.org/10.48550/arXiv.2012.07581 arXiv:2012.07581 [cs].

[6] Yoana Ahmetoglu, Duncan P. Brumby, and Anna L. Cox. 2021. To Plan or Not to Plan? A Mixed-Methods Diary Study Examining When, How and Why Knowledge Work Planning is Inaccurate. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 222:1–222:20. https://doi.org/10.1145/3432921

[7] Adam L. Alter, Daniel M. Oppenheimer, Nicholas Epley, and Rebecca N. Eyre. 2007. Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology. General* 136, 4 (Nov. 2007), 569–576. https://doi.org/10.1037/0096-3445.136.4.569

[8] Amine Amzil. 2013. The Effect of a Metacognitive Intervention on College Students' Reading Performance and Metacognitive Skills. *Journal of Educational and Developmental Psychology* 4, 1 (Dec. 2013), p27. https://doi.org/10.5539/jedp.v4n1p27

[9] Heidi Goodrich Andrade. 2000. Using Rubrics To Promote Thinking and Learning. *Educational Leadership* 57, 5 (2000), 13–18. ERIC Number: EJ609600.

[10] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. https://doi.org/10.48550/arXiv.1909.03012 arXiv:1909.03012 [cs, stat].

[11] Roger Azevedo. 2020. Reflections on the field of metacognition: issues, challenges, and opportunities. *Metacognition and Learning* 15, 2 (Aug. 2020), 91–98. https://doi.org/10.1007/s11409-020-09231-x

[12] Lisanne Bainbridge. 1983. Ironies of automation. *Automatica* 19, 6 (Nov. 1983), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

[13] Albert Bandura. 1997. *Self-efficacy: The exercise of control.* W H Freeman/Times Books/ Henry Holt & Co, New York, NY, US. Pages: ix, 604.

[14] Shraddha Barke, Michael B. James, and Nadia Polikarpova. 2023. Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proceedings of the ACM on Programming Languages* 7, OOPSLA1 (April 2023), 78:85–78:111. https://doi.org/10.1145/3586030

[15] Frederic Becker, Maria Wirzberger, Viktoria Pammer-Schindler, Srinidhi Srinivas, and Falk Lieder. 2023. Systematic metacognitive reflection helps people discover far-sighted decision strategies: A process-tracing experiment. *Judgment and Decision Making* 18 (Jan. 2023), e15. https://doi.org/10.1017/jdm.2023.16 Publisher: Cambridge University Press.

[16] Aaron S. Benjamin, Robert A. Bjork, and Bennett L. Schwartz. 1998. The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General* 127, 1 (1998), 55–68. https://doi.org/10.1037/0096-3445.127.1.55 Place: US Publisher: American Psychological Association.

[17] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (March 2022), 2:1–2:27. https://doi.org/10.1145/3517233

[18] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581314

[19] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571

[20] A.F. Blackwell. 2002. First steps in programming: a rationale for attention investment models. In *Proceedings IEEE 2002 Symposia on Human Centric Computing Languages and Environments*. 2–10. https://doi.org/10.1109/HCC.2002.1046334

[21] Monique Boekaerts and Lyn Corno. 2005. Self-Regulation in the Classroom: A Perspective on Assessment and Intervention. *Applied Psychology: An International Review* 54, 2 (2005), 199–231. https://doi.org/10.1111/j.1464-0597.2005.00205.x Place: United Kingdom Publisher: Blackwell Publishing.

[22] Annika Boldt and Sam J. Gilbert. 2019. Confidence guides spontaneous cognitive offloading. *Cognitive Research: Principles and Implications* 4, 1 (Dec. 2019), 45. https://doi.org/10.1186/s41235-019-0195-y

[23] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 arXiv:2108.07258 [cs].

[24] Mimi Bong and Einar M. Skaalvik. 2003. Academic Self-Concept and Self-Efficacy: How Different Are They Really? *Educational Psychology Review* 15, 1 (March 2003), 1–40. https://doi.org/10.1023/A:1021302408382

[25] Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M. Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, Christoph Miksovic, Paolo Scotton, Kartik Talamadupula, and Gegi Thomas. 2023. Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 220–239. https://doi.org/10.1145/3581641.3584088

[26] Daniel Buschek, Malin Eiband, and Heinrich Hussmann. 2022. How to Support Users in Understanding Intelligent Systems? An Analysis and Conceptual Framework of User Questions Considering User Mindsets, Involvement, and Knowledge Outcomes. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (Nov. 2022), 29:1–29:27. https://doi.org/10.1145/3519264

[27] Deborah L. Butler. 1998. The strategic content learning approach to promoting self-regulated learning: A report of three studies. *Journal of Educational Psychology* 90, 4 (1998), 682–697. https://doi.org/10.1037/0022-0663.90.4.682 Place:

US Publisher: American Psychological Association.

[28] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[29] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. 2019. Better beware: comparing metacognition for phishing and legitimate emails. *Metacognition and Learning* 14, 3 (Dec. 2019), 343–362. https://doi.org/10.1007/s11409-019-09197-5

[30] Jason Carpenter, Maxine T. Sherman, Rogier A. Kievit, Anil K. Seth, Hakwan Lau, and Stephen M. Fleming. 2019. Domain-General Enhancements of Metacognitive Ability Through Adaptive Training. *Journal of Experimental Psychology. General* 148, 1 (Jan. 2019), 51–64. https://doi.org/10.1037/xge0000505

[31] Cameron S. Carter and Vincent van Veen. 2007. Anterior cingulate cortex and conflict detection: An update of theory and data. *Cognitive, Affective, & Behavioral Neuroscience* 7, 4 (Dec. 2007), 367–379. https://doi.org/10.3758/CABN.7.4.367

[32] J. J. Cañas *, A. AntolÍ, I. Fajardo, and L. Salmerón. 2005. Cognitive inflexibility and the development and use of strategies for solving complex dynamic problems: effects of different types of training. *Theoretical Issues in Ergonomics Science* 6, 1 (Jan. 2005), 95–108. https://doi.org/10.1080/14639220512331311599 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14639220512331311599.

[33] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. https://doi.org/10.48550/arXiv.2202.04092 arXiv:2202.04092 [cs].

[34] Xiang 'Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K. Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl D. D. Willis, Chien-Sheng Wu, and Bolei Zhou. 2023. Next Steps for Human-Centered Generative AI: A Technical Perspective. https://doi.org/10.48550/arXiv.2306.15774 arXiv:2306.15774 [cs].

[35] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (Feb. 2022), 107018. https://doi.org/10.1016/j.chb.2021.107018

[36] David Church and Mark Carroll. 2023. How does metacognition improve decision-making in healthcare practitioners? *Journal of Paramedic Practice* 15, 3 (March 2023), 113–123. https://doi.org/10.12968/jpar.2023.15.3.113 Publisher: Mark Allen Group.

[37] Timothy J. Cleary and Barry J. Zimmerman. 2004. Self-Regulation Empowerment Program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools* 41, 5 (May 2004), 537–550. https://doi.org/10.1002/pits.10177 Publisher: Wiley-Liss Inc..

[38] Anita Crescenzi, Austin R. Ward, Yuan Li, and Rob Capra. 2021. Supporting Metacognition during Exploratory Search with the OrgBox. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1197–1207. https://doi.org/10.1145/3404835.3462955

[39] David R. Cross and Scott G. Paris. 1988. Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology* 80, 2 (1988), 131–142. https://doi.org/10.1037/0022-0663.80.2.131 Place: US Publisher: American Psychological Association.

[40] Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C. Desmarais, Zhen Ming, and Jiang. 2023. GitHub Copilot AI pair programmer: Asset or Liability? https://doi.org/10.48550/arXiv.2206.15331 arXiv:2206.15331 [cs].

[41] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3580969

[42] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3502141

[43] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. https://doi.org/10.48550/arXiv.2209.01390 arXiv:2209.01390 [cs].

[44] Sophie De Beukelaer, Neza Vehar, Max Rollwage, Stephen M. Fleming, and Manos Tsakiris. 2023. Changing minds about climate change: a pervasive role for domain-general metacognition. *Humanities and Social Sciences Communications* 10, 1 (Feb. 2023), 1–10. https://doi.org/10.1057/s41599-023-01528-x Number: 1 Publisher: Palgrave.

[45] Hester de Boer, Anouk S. Donker, Danny D. N. M. Kostons, and Greetje P. C. van der Werf. 2018. Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review* 24 (June 2018), 98–115. https://doi.org/10.1016/j.edurev.2018.03.002

[46] Anique B. H. de Bruin, Julian Roelle, Shana K. Carpenter, Martine Baars, and EFG-MRE. 2020. Synthesizing Cognitive Load and Self-regulation Theory: a Theoretical Framework and Research Agenda. *Educational Psychology Review* 32, 4 (Dec. 2020), 903–915. https://doi.org/10.1007/s10648-020-09576-4

[47] Peter de Vries, Cees Midden, and Don Bouwhuis. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies* 58, 6 (June 2003), 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9

[48] Victor R. Delclos and Christine Harrington. 1991. Effects of strategy monitoring and proactive instruction on children's problem-solving performance. *Journal of Educational Psychology* 83, 1 (1991), 35–42. https://doi.org/10.1037/0022-0663.83.1.35 Place: US Publisher: American Psychological Association.

[49] Kobe Desender, Annika Boldt, and Nick Yeung. 2018. Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science* 29, 5 (May 2018), 761–778. https://doi.org/10.1177/0956797617744771 Publisher: SAGE Publications Inc.

[50] Design Lab. 2017. What does it mean to be literate in the Age of Google? | Dan Russell | Design@Large. https://www.youtube.com/watch?v=SgOBrYOttZg

[51] Anneline Devolder, Johan van Braak, and Jo Tondeur. 2012. Supporting self-regulated learning in computer-based learning environments: Systematic review of effects of scaffolding in the domain of science education. *Journal of Computer Assisted Learning* 28 (Dec. 2012). https://doi.org/10.1111/j.1365-2729.2011.00476.x

[52] A. S. Donker, H. de Boer, D. Kostons, C. C. Dignath van Ewijk, and M. P. C. van der Werf. 2014. Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review* 11 (Jan. 2014), 1–26. https://doi.org/10.1016/j.edurev.2013.11.002

[53] Timothy L. Dunn, Connor Gaspar, Daev McLean, Derek J. Koehler, and Evan F. Risko. 2021. Distributed metacognition: Increased bias and deficits in metacognitive sensitivity when retrieving information from the internet. *Technology, Mind, and Behavior* 2, 3 (Aug. 2021). https://doi.org/10.1037/tmb0000039

[54] Jacquelynne S. Eccles and Allan Wigfield. 2002. Motivational Beliefs, Values, and Goals. *Annual Review of Psychology* 53, 1 (2002), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153 _eprint: https://doi.org/10.1146/annurev.psych.53.100901.135153.

[55] Anastasia Efklides. 2008. Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist* 13, 4 (2008), 277–287. https://doi.org/10.1027/1016-9040.13.4.277 Place: Germany Publisher: Hogrefe & Huber Publishers.

[56] Anastasia Efklides and Plousia Misailidi (Eds.). 2010. *Trends and Prospects in Metacognition Research*. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4419-6546-2

[57] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence (Lecture Notes in Computer Science)*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

[58] Emmaline Drew Eliseev and Elizabeth J. Marsh. 2023. Understanding why searching the internet inflates confidence in explanatory ability. *Applied Cognitive Psychology* 37, 4 (2023), 711–720. https://doi.org/10.1002/acp.4058 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.4058.

[59] K.A. Ericsson and H.A. Simon. 1993. *Protocol Analysis*. The MIT Press. https://mitpress.mit.edu/9780262550239/protocol-analysis/

[60] Anneli Eteläpelto. 1993. Metacognition and the Expertise of Computer Program Comprehension. *Scandinavian Journal of Educational Research* 37, 3 (Jan. 1993), 243–254. https://doi.org/10.1080/0031383930370305 Publisher: Routledge _eprint: https://doi.org/10.1080/0031383930370305.

[61] Kasra Ferdowsifard, Allen Ordookhanians, Hila Peleg, Sorin Lerner, and Nadia Polikarpova. 2020. Small-Step Live Programming by Example. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 614–626. https://doi.org/10.1145/3379337.3415869

[62] Klaus Fiedler, Rakefet Ackerman, and Chiara Scarampi. 2019. Metacognition: Monitoring and controlling one's own knowledge, reasoning and decisions. *The psychology of human thought: An introduction* (2019), 89–111. Publisher: Heidelberg University Publishing: Heidelberg.

[63] Matthew Fisher and Daniel M. Oppenheimer. 2021. Harder Than You Think: How Outside Assistance Leads to Overconfidence. *Psychological Science* 32, 4 (April 2021), 598–610. https://doi.org/10.1177/0956797620975779

[64] John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist* 34, 10 (1979), 906–911. https://doi.org/10.1037/0003-066X.34.10.906 Place: US Publisher: American Psychological Association.

[65] Stephen Fleming. 2023. Metacognition and confidence: A review and synthesis. https://doi.org/10.31234/osf.io/ge7tz

[66] Stephen M. Fleming and Hakwan C. Lau. 2014. How to measure metacognition. *Frontiers in Human Neuroscience* 8 (2014). https://www.frontiersin.org/articles/10.3389/fnhum.2014.00443

[67] Petros Georghiades. 2004. From the general to the situated: three decades of metacognition. *International Journal of Science Education* 26, 3 (Feb. 2004), 365–383. https://doi.org/10.1080/0950069032000119401 Publisher: Routledge _eprint: https://doi.org/10.1080/0950069032000119401.

[68] Sam J. Gilbert. 2015. Strategic offloading of delayed intentions into the external environment. *Quarterly Journal of Experimental Psychology* 68, 5 (May 2015), 971–992. https://doi.org/10.1080/17470218.2014.972963 Publisher: SAGE Publications.

[69] Sam J. Gilbert, Annika Boldt, Chhavi Sachdeva, Chiara Scarampi, and Pei-Chun Tsai. 2023. Outsourcing Memory to External Tools: A Review of 'Intention Offloading'. *Psychonomic Bulletin & Review* 30, 1 (Feb. 2023), 60–76. https://doi.org/10.3758/s13423-022-02139-4

[70] Frederic Gmeiner, Humphrey Yang, Lining Yao, Kenneth Holstein, and Nikolas Martelaro. 2023. Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3580999

[71] Andrew D. Gordon, Carina Negreanu, José Cambronero, Rasika Chakravarthy, Ian Drosos, Hao Fang, Bhaskar Mitra, Hannah Richardson, Advait Sarkar, Stephanie Simmons, Jack Williams, and Ben Zorn. 2023. *Co-audit for copilots: tools to help humans double-check AI-generated content*. Technical Report. Microsoft Research. https://aka.ms/co-audit

[72] Anthony M. Grant, John Franklin, and Peter Langford. 2002. The Self-Reflection and Insight Scale: A new measure of private self-consciousness. *Social Behavior and Personality: An International Journal* 30, 8 (2002), 821–835. https://doi.org/10.2224/sbp.2002.30.8.821 Place: New Zealand Publisher: Society for Personality Research.

[73] Jane Gravill, Deborah Compeau, and Barbara Marcolin. 2002. Metacognition and IT: The influence of self-efficacy and self-awareness. In *AMCIS 2002 Proceedings*. 147.

[74] Ken Gu, Madeleine Grunde-McLaughlin, Andrew M. McNutt, Jeffrey Heer, and Tim Althoff. 2023. How Do Data Analysts Respond to AI Assistance? A Wizard-of-Oz Study. https://doi.org/10.48550/arXiv.2309.10108 arXiv:2309.10108 [cs].

[75] Jo E. Hannay, Tore Dybå, Erik Arisholm, and Dag I. K. Sjøberg. 2009. The effectiveness of pair programming: A meta-analysis. *Information and Software Technology* 51, 7 (July 2009), 1110–1122. https://doi.org/10.1016/j.infsof.2009.02.001

[76] Neville Hatton and David Smith. 1995. Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education* 11, 1 (Jan. 1995), 33–49. https://doi.org/10.1016/0742-051X(94)00012-U

[77] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581025

[78] Jake M. Hofman, Daniel G. Goldstein, and David M. Rothschild. 2023. A Sports Analogy for Understanding Different Ways to Use AI. *Harvard Business Review* (Dec. 2023). https://hbr.org/2023/12/a-sports-analogy-for-understanding-different-ways-to-use-ai Section: AI and machine learning.

[79] Xiao Hu, Liang Luo, and Stephen M. Fleming. 2019. A role for metamemory in cognitive offloading. *Cognition* 193 (Dec. 2019), 104012. https://doi.org/10.1016/j.cognition.2019.104012

[80] Jessica D. Huff and John L. Nietfeld. 2009. Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning* 4, 2 (2009), 161–176. https://doi.org/10.1007/s11409-009-9042-8 Place: Germany Publisher: Springer.

[81] Sarah Inman and David Ribes. 2019. "Beautiful Seams": Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300508

[82] Christian P. Janssen, Stella F. Donker, Duncan P. Brumby, and Andrew L. Kun. 2019. History and future of human-automation interaction. *International Journal of Human-Computer Studies* 131 (Nov. 2019), 99–107. https://doi.org/10.1016/j.ijhcs.2019.05.006

[83] Dhanya Jayagopal, Justin Lubin, and Sarah E. Chasins. 2022. Exploring the Learnability of Program Synthesizers by Novice Programmers. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3526113.3545659

[84] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. https://doi.org/10.1145/3571730 arXiv:2202.03629 [cs].

[85] Ellen Jiang, Edwin Toh, Alejandra Molina, Kristen Olson, Claire Kayacik, Aaron Donsbach, Carrie J Cai, and Michael Terry. 2022. Discovering the Syntax and Strategies of Natural Language Programming with Generative Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3501870

[86] David H. Jonassen. 1997. Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educational Technology Research and Development* 45, 1 (March 1997), 65–94. https://doi.org/10.1007/BF02299613

[87] Sucharit Katyal and Stephen Fleming. 2023. Construct validity in metacognition research: balancing the tightrope between rigor of measurement and breadth of construct. https://doi.org/10.31234/osf.io/etjqh

[88] Alan Kay. 1990. User Interface: A Personal View. In *The Art of Human-Computer Interface Design*. 191–207. http://ui.korea.ac.kr/Board/Upload/a%20personal%20view_n.pdf

[89] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J. Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3544548.3580919

[90] Nina Keith and Michael Frese. 2005. Self-Regulation in Error Management Training: Emotion Control and Metacognition as Mediators of Performance Effects. *Journal of Applied Psychology* 90, 4 (2005), 677–691. https://doi.org/10.1037/0021-9010.90.4.677 Place: US Publisher: American Psychological Association.

[91] Alison King. 1991. Improving lecture comprehension: Effects of a metacognitive strategy. *Applied Cognitive Psychology* 5, 4 (1991), 331–346. https://doi.org/10.1002/acp.2350050404 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050404.

[92] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (July 2018), 1–26. https://doi.org/10.1145/3214273

[93] Asher Koriat. 2007. Metacognition and consciousness. In *The Cambridge handbook of consciousness*. Cambridge University Press, New York, NY, US, 289–325. https://doi.org/10.1017/CBO9780511816789.012

[94] Asher Koriat and Robert A. Bjork. 2005. Illusions of Competence in Monitoring One's Knowledge During Study. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 2 (2005), 187–194. https://doi.org/10.1037/0278-7393.31.2.187 Place: US Publisher: American Psychological Association.

[95] Asher Koriat, Hilit Ma'ayan, and Ravit Nussinson. 2006. The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General* 135, 1 (Feb. 2006), 36–69. https://doi.org/10.1037/0096-3445.135.1.36

[96] Bracha Kramarski and Zemira R. Mevarech. 2003. Enhancing Mathematical Reasoning in the Classroom: The Effects of Cooperative Learning and Metacognitive Training. *American Educational Research Journal* 40, 1 (2003), 281–310. https://www.jstor.org/stable/3699433 Publisher: [American Educational Research Association, Sage Publications, Inc.].

[97] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678

[98] Steinar Kvale. 1994. *InterViews: An introduction to qualitative research interviewing*. Sage Publications, Inc, Thousand Oaks, CA, US. Pages: xvii, 326.

[99] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[100] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. https://doi.org/10.48550/arXiv.2202.01875 arXiv:2202.01875 [cs].

[101] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (July 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[102] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies* 40, 1 (Jan. 1994), 153–184. https://doi.org/10.1006/ijhc.1994.1007

[103] Jenny T. Liang, Chenyang Yang, and Brad A. Myers. 2023. Understanding the Usability of AI Programming Assistants. https://doi.org/10.48550/arXiv.2303.17125 arXiv:2303.17125 [cs].

[104] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. https://doi.org/10.48550/arXiv.2306.01941 arXiv:2306.01941 [cs].

[105] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–31. https://doi.org/10.1145/3544548.3580817

[106] Jennifer A. Livingston. 2003. *Metacognition: An Overview*. Technical Report. https://eric.ed.gov/?id=ED474273 ERIC Number: ED474273.

[107] Dastyni Loksa, Lauren Margulieux, Brett A. Becker, Michelle Craig, Paul Denny, Raymond Pettit, and James Prather. 2022. Metacognition and Self-Regulation in Programming Education: Theories and Exemplars of Use. *ACM Transactions on Computing Education* 22, 4 (Sept. 2022), 39:1–39:31. https://doi.org/10.1145/3487050

[108] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445562

[109] Omar López-Vargas, Jaime Ibáñez-Ibáñez, and Oswaldo Racines-Prada. 2017. Students' Metacognition and Cognitive Style and Their Effect on Cognitive Load and Learning Achievement. *Journal of Educational Technology & Society* 20, 3 (2017), 145–157. https://www.jstor.org/stable/26196126 Publisher: International Forum of Educational Technology & Society.

[110] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581058

[111] Tadhg E. MacIntyre, Eric R. Igou, Mark J. Campbell, Aidan P. Moran, and James Matthews. 2014. Metacognition and action: a new pathway to understanding social and cognitive aspects of expertise in sport. *Frontiers in Psychology* 5 (2014). https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01155

[112] Thomas P. Mackey and Trudi E. Jacobson. 2017. Reframing Information Literacy as a Metaliteracy | Mackey | College & Research Libraries. (April 2017). https://doi.org/10.5860/crl-76r1

[113] Stephen MacNeil, Andrew Tran, Joanne Kim, Ziheng Huang, Seth Bernstein, and Dan Mogil. 2023. Prompt Middleware: Mapping Prompts for Large Language Models to UI Affordances. https://doi.org/10.48550/arXiv.2307.01142 arXiv:2307.01142 [cs].

[114] P. Madhavan and D. A. Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (July 2007), 277–301. https://doi.org/10.1080/14639220500337708 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14639220500337708.

[115] Brian Maniscalco and Hakwan Lau. 2014. Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance SDT model. In *The cognitive neuroscience of metacognition*. Springer-Verlag Publishing, New York, NY, US, 25–66. https://doi.org/10.1007/978-3-642-45190-4_3

[116] Gennie Mansi and Mark Riedl. 2023. Why Don't You Do Something About It? Outlining Connections between AI Explanations and User Actions. http://arxiv.org/abs/2305.06297 arXiv:2305.06297 [cs].

[117] Matthew M. Martin and Rebecca B. Rubin. 1995. A New Measure of Cognitive Flexibility. *Psychological Reports* 76, 2 (April 1995), 623–626. https://doi.org/10.2466/pr0.1995.76.2.623 Publisher: SAGE Publications Inc.

[118] Audrey Mazancieux, Michael Pereira, Nathan Faivre, Pascal Mamassian, Chris J. A. Moulin, and Céline Souchay. 2023. Towards a common conceptual space for metacognition in perception and memory. *Nature Reviews Psychology* (Nov. 2023), 1–16. https://doi.org/10.1038/s44159-023-00245-1 Publisher: Nature Publishing Group.

[119] Andrew M Mcnutt, Chenglong Wang, Robert A Deline, and Steven M. Drucker. 2023. On the Design of AI-powered Code Assistants for Notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3580940

[120] Lynn Meltzer. 2014. Teaching Executive Functioning Processes: Promoting Metacognition, Strategy Use, and Effort. In *Handbook of Executive Functioning*, Sam Goldstein and Jack A. Naglieri (Eds.). Springer, New York, NY, 445–473. https://doi.org/10.1007/978-1-4614-8106-5_25

[121] Zemira R. Mevarech and Chagit Amrany. 2008. Immediate and delayed effects of meta-cognitive instruction on regulation of cognition and mathematics achievement. *Metacognition and Learning* 3, 2 (Aug. 2008), 147–157. https://doi.org/10.1007/s11409-008-9023-3

[122] André N. Meyer, Gail C. Murphy, Thomas Zimmermann, and Thomas Fritz. 2021. Enabling Good Work Habits in Software Developers through Reflective Goal-Setting. *IEEE Transactions on Software Engineering* 47, 9 (Sept. 2021), 1872–1885. https://doi.org/10.1109/TSE.2019.2938525 Conference Name: IEEE Transactions on Software Engineering.

[123] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[124] Stephen Monsell. 2003. Task switching. *Trends in Cognitive Sciences* 7, 3 (March 2003), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

[125] J. Moon. 2000. Learning Journals | A Handbook for Reflective Practice and Professiona. https://www.taylorfrancis.com/books/mono/10.4324/9780203969212/learning-journals-jennifer-moon

[126] Daniel Muijs and Christian Bokhove. 2020. *Metacognition and Self-Regulation: Evidence Review*. Technical Report. Education Endowment Foundation. https://eric.ed.gov/?id=ED612286 Publication Title: Education Endowment Foundation ERIC Number: ED612286.

[127] T. O. Nelson. 1984. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin* 95, 1 (Jan. 1984), 109–133.

[128] Thomas O. Nelson. 1990. Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation*, Gordon H. Bower (Ed.). Vol. 26. Academic Press, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

[129] Thomas O. Nelson and John Dunlosky. 1991. When People's Judgments of Learning (JOLs) are Extremely Accurate at Predicting Subsequent Recall: The "Delayed-JOL Effect". *Psychological Science* 2, 4 (July 1991), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x Publisher: SAGE Publications Inc.

[130] Ikujiro Nonaka and Ryoko Toyama. 2003. The knowledge-creating theory revisited: knowledge creation as a synthesizing process. *Knowledge Management Research & Practice* 1, 1 (July 2003), 2–10. https://doi.org/10.1057/palgrave.kmrp.8500001 Publisher: Taylor & Francis _eprint: https://doi.org/10.1057/palgrave.kmrp.8500001.

[131] Elisabeth Norman, Gerit Pfuhl, Rannveig Grøm Sæle, Frode Svartdal, Torstein Låg, and Tove Irene Dahl. 2019. Metacognition in Psychology. *Review of General Psychology* 23, 4 (Dec. 2019), 403–424. https://doi.org/10.1177/1089268019883821 Publisher: SAGE Publications Inc.

[132] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (July 2023), 187–192. https://doi.org/10.1126/science.adh2586 Publisher: American Association for the Advancement of Science.

[133] Matthias Nückles, Sandra Hübner, Sandra Dümer, and Alexander Renkl. 2010. Expertise reversal effects in writing-to-learn. *Instructional Science* 38, 3 (May 2010), 237–258. https://doi.org/10.1007/s11251-009-9106-9

[134] Matthias Nückles, Julian Roelle, Inga Glogger-Frey, Julia Waldeyer, and Alexander Renkl. 2020. The Self-Regulation-View in Writing-to-Learn: Using Journal Writing to Optimize Cognitive Load in Self-Regulated Learning. *Educational Psychology Review* 32, 4 (Dec. 2020), 1089–1126. https://doi.org/10.1007/s10648-020-09541-1

[135] OpenAI. 2023. Customizing GPT-3 for your application. https://openai.com/blog/customizing-gpt-3

[136] Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2023. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. http://arxiv.org/abs/2308.02828 arXiv:2308.02828 [cs].

[137] Annemarie Sullivan Palincsar and Ann L. Brown. 1984. Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities. *Cognition and Instruction* 1, 2 (1984), 117–175. https://www.jstor.org/stable/3233567 Publisher: Taylor & Francis, Ltd..

[138] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 102:1–102:15. https://doi.org/10.1145/3359204

[139] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2022. Do Users Write More Insecure Code with AI Assistants? https://doi.org/10.48550/arXiv.2211.03622 arXiv:2211.03622 [cs].

[140] Nancy E. Perry, Lynda Hutchinson, and Carolyn Thauberger. 2008. Talking about Teaching Self-Regulated Learning: Scaffolding Student Teachers' Development and Use of Practices that Promote Self-Regulated Learning. *International Journal of Educational Research* 47, 2 (2008), 97–108. https://doi.org/10.1016/j.ijer.2007.11.010 Publisher: Elsevier ERIC Number: EJ798056.

[141] Peter Pirolli and Stuart Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. 2–4.

[142] James Prather, Brett A. Becker, Michelle Craig, Paul Denny, Dastyni Loksa, and Lauren Margulieux. 2020. What Do We Think We Think We Are Doing?: Metacognition and Self-Regulation in Programming. In *Proceedings of the 2020 ACM Conference on International Computing Education Research*. ACM, Virtual Event New Zealand, 2–13. https://doi.org/10.1145/3372782.3406263

[143] James Prather, Brent N. Reeves, Paul Denny, Brett A. Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2023. "It's Weird That it Knows What I Want": Usability and Interactions with Copilot for Novice Programmers. https://doi.org/10.48550/arXiv.2304.02491 arXiv:2304.02491 [cs].

[144] David D. Preiss. 2022. Metacognition, Mind Wandering, and Cognitive Flexibility: Understanding Creativity. *Journal of Intelligence* 10, 3 (Sept. 2022), 69. https://doi.org/10.3390/jintelligence10030069 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[145] Dobromir Rahnev. 2023. Measuring metacognition: A comprehensive assessment of current methods. https://doi.org/10.31234/osf.io/waz9h

[146] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. https://doi.org/10.48550/arXiv.2304.09991 arXiv:2304.09991 [cs].

[147] Leon Reicherts and Yvonne Rogers. 2020. Do Make me Think! How CUIs Can Support Cognitive Processes. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3405755.3406157

[148] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20, 9 (Sept. 2016), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[149] Evan F. Risko and Megan O. Kelly. 2023. Thinking in the digital age: Everyday cognition and the dawn of a new age of metacognition research. *Applied Cognitive Psychology* 37, 4 (2023), 785–788. https://doi.org/10.1002/acp.4102 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.4102.

[150] Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D. Weisz. 2023. The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 491–514. https://doi.org/10.1145/3581641.3584037

[151] Marion Rouault, Andrew McWilliams, Micah G. Allen, and Stephen M. Fleming. 2018. Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience* 1 (2018), e17. https://doi.org/10.1017/pen.2018.16

[152] Martin Ruskov. 2023. Grimm in Wonderland: Prompt Engineering with Midjourney to Illustrate Fairytales. https://doi.org/10.48550/arXiv.2302.08961 arXiv:2302.08961 [cs].

[153] Nikita A. Salovich and David N. Rapp. 2021. Misinformed and unaware? Metacognition and the influence of inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47, 4 (2021), 608–624. https://doi.org/10.1037/xlm0000977 Place: US Publisher: American Psychological Association.

[154] Advait Sarkar. 2023. Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3596671.3597650

[155] Advait Sarkar. 2023. Should Computers Be Easy To Use? Questioning the Doctrine of Simplicity in User Interface Design. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3544549.3582741

[156] Advait Sarkar, Andrew D. Gordon, Carina Negreanu, Christian Poelitz, Sruti Srinivasa Ragavan, and Ben Zorn. 2022. What is it like to program with artificial intelligence? https://doi.org/10.48550/arXiv.2208.06213 arXiv:2208.06213 [cs].

[157] Wout Schellaert, Fernando Martínez-Plumed, Karina Vold, John Burden, Pablo A. M. Casares, Bao Sheng Loe, Roi Reichart, Sean Ô hÉigeartaigh, Anna Korhonen, and José Hernández-Orallo. 2023. Your Prompt is My Command: On Assessing the Human-Centred Generality of Multimodal Models. *Journal of Artificial Intelligence Research* 77 (June 2023), 377–394. https://doi.org/10.1613/jair.1.14157

[158] Lion Schulz, Max Rollwage, Raymond J. Dolan, and Stephen M. Fleming. 2020. Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences* 117, 49 (Dec. 2020), 31527–31534. https://doi.org/10.1073/pnas.2009641117 Publisher: Proceedings of the National Academy of Sciences.

[159] Dale H. Schunk and Peggy A. Ertmer. 2000. Self-regulation and academic learning: Self-efficacy enhancing interventions. In *Handbook of self-regulation*. Academic Press, San Diego, CA, US, 631–649. https://doi.org/10.1016/B978-012109890-2/50048-2

[160] Rolf Schwonke. 2015. Metacognitive Load – Useful, or Extraneous Concept? Metacognitive and Self-Regulatory Demands in Computer-Based Learning. *Journal of Educational Technology & Society* 18, 4 (2015), 172–184. https://www.jstor.org/stable/jeductechsoci.18.4.172 Publisher: International Forum of Educational Technology & Society.

[161] Ava Scott and Sam Gilbert. 2023. Metacognition guides intention offloading and fulfilment of real-world plans. https://doi.org/10.31234/osf.io/y46mq

[162] James R. Segedy, John S. Kinnebrew, Benjamin S. Goldberg, Robert A. Sottilare, and Gautam Biswas. 2015. Designing Representations and Support for Metacognition in the Generalized Intelligent Framework for Tutoring. In *Foundations of Augmented Cognition (Lecture Notes in Computer Science)*, Dylan D. Schmorrow and Cali M. Fidopiastis (Eds.). Springer International Publishing, Cham, 663–674. https://doi.org/10.1007/978-3-319-20816-9_63

[163] Tina Seufert. 2018. The interplay between self-regulation in learning and cognitive load. *Educational Research Review* 24 (June 2018), 116–129. https://doi.org/10.1016/j.edurev.2018.03.004

[164] Preya Shabrina, Behrooz Mostafavi, Mark Abdelshiheed, Min Chi, and Tiffany Barnes. 2023. Investigating the Impact of Backward Strategy Learning in a Logic Tutor: Aiding Subgoal Learning Towards Improved Problem Solving. *International Journal of Artificial Intelligence in Education* (Aug. 2023). https://doi.org/10.1007/s40593-023-00338-1

[165] Auste Simkute, Ewa Luger, Mike Evans, and Rhianne Jones. 2020. Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference (DIS' 20 Companion)*. Association for Computing Machinery, New York, NY, USA, 263–268. https://doi.org/10.1145/3393914.3395862

[166] Auste Simkute, Lev Tankelevitch, Viktor Kewenig, Ava Elizabeth Scott, Abigail Sellen, and Sean Rintel. 2024. Ironies of Generative AI: Understanding and mitigating productivity loss in human-AI interactions. https://doi.org/10.48550/arXiv.2402.11364 arXiv:2402.11364 [cs].

[167] Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. http://arxiv.org/abs/2307.03744 arXiv:2307.03744 [cs].

[168] Rand J. Spiro, Paul J. Feltovich, Paul L. Feltovich, Michael J. Jacobson, and Richard L. Coulson. 1991. Cognitive Flexibility, Constructivism, and Hypertext: Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains. *Educational Technology* 31, 5 (1991), 24–33. https://www.jstor.org/stable/44427517 Publisher: Educational Technology Publications, Inc..

[169] Sruti Srinivasa Ragavan, Zhitao Hou, Yun Wang, Andrew D Gordon, Haidong Zhang, and Dongmei Zhang. 2022. GridBook: Natural Language Formulas for the Spreadsheet Grid. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 345–368. https://doi.org/10.1145/3490099.3511161

[170] Keith E. Stanovich and Richard F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* 23, 5 (2000), 645–665. https://doi.org/10.1017/S0140525X00003435 Place: United Kingdom Publisher: Cambridge University Press.

[171] Mark Steyvers and Aakriti Kumar. 2023. Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* (July 2023), 17456916231181102. https://doi.org/10.1177/17456916231181102

[172] Sean M. Stone and Benjamin C. Storm. 2021. Search fluency as a misleading measure of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47, 1 (2021), 53–64. https://doi.org/10.1037/xlm0000806 Place: US Publisher: American Psychological Association.

[173] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. https://doi.org/10.48550/arXiv.2305.11483 arXiv:2305.11483 [cs].

[174] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 212–228. https://doi.org/10.1145/3490099.3511119

[175] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445088

[176] Adam Svendsen and Bruce Garvey. 2023. An Outline for an Interrogative/Prompt Library to help improve output quality from Generative-AI Datasets. https://doi.org/10.2139/ssrn.4495319

[177] John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* 12, 2 (1988), 257–285. https://doi.org/10.1207/s15516709cog1202_4 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1202_4.

[178] John Sweller, Jeroen J. G. van Merrienboer, and Fred G. W. C. Paas. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10, 3 (Sept. 1998), 251–296. https://doi.org/10.1023/A:1022193728205

[179] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2022. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. https://doi.org/10.48550/arXiv.2210.04885 arXiv:2210.04885 [cs].

[180] Kimberly D. Tanner. 2012. Promoting Student Metacognition. *CBE—Life Sciences Education* 11, 2 (June 2012), 113–120. https://doi.org/10.1187/cbe.12-03-0033 Publisher: American Society for Cell Biology (lse).

[181] Pina Tarricone. 2011. *The Taxonomy of Metacognition.* Psychology Press. Google-Books-ID: c1p6AgAAQBAJ.

[182] Heliodoro Tejeda, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2022. AI-Assisted Decision-making: a Cognitive Modeling Approach to Infer Latent Reliance Strategies. *Computational Brain & Behavior* 5, 4 (Dec. 2022), 491–508. https://doi.org/10.1007/s42113-022-00157-y

[183] Valerie Thompson and Kinga Morsanyi. 2012. Analytic thinking: do you feel like it? *Mind & Society* 11, 1 (June 2012), 93–105. https://doi.org/10.1007/s11299-012-0100-6

[184] Valerie A. Thompson, Jamie A. Prowse Turner, Gordon Pennycook, Linden J. Ball, Hannah Brack, Yael Ophir, and Rakefet Ackerman. 2013. The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition* 128, 2 (Aug. 2013), 237–251. https://doi.org/10.1016/j.cognition.2012.09.012

[185] Sascha Topolinski and Rolf Reber. 2010. Immediate truth – Temporal contiguity between a cognitive problem and its solution determines experienced veracity of the solution. *Cognition* 114, 1 (Jan. 2010), 117–122. https://doi.org/10.1016/j.cognition.2009.09.009

[186] A. K. Troyer, M. Moscovitch, and G. Winocur. 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology* 11, 1 (Jan. 1997), 138–146. https://doi.org/10.1037//0894-4105.11.1.138

[187] Christian Unkelbach and Rainer Greifeneder. 2013. A general model of fluency effects in judgment and decision making. In *The experience of thinking: How the fluency of mental processes influences cognition and behaviour.* Psychology Press, New York, NY, US, 11–32.

[188] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22).* Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3491101.3519665

[189] Martin Valcke. 2002. Cognitive load: updating the theory? *Learning and Instruction* (2002).

[190] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. https://doi.org/10.48550/arXiv.2302.07248 arXiv:2302.07248 [cs].

[191] Jo Vermeulen, Kris Luyten, Elise van den Hoven, and Karin Coninx. 2013. Crossing the bridge over Norman's Gulf of Execution: revealing feedforward's true identity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* Association for Computing Machinery, New York, NY, USA, 1931–1940. https://doi.org/10.1145/2470654.2466255

[192] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[193] Tingting Wang, Shan Li, Xiaoshan Huang, Zexuan Pan, and Susanne P. Lajoie. 2023. Examining students' cognitive load in the context of self-regulated learning with an intelligent tutoring system. *Education and Information Technologies* 28, 5 (May 2023), 5697–5715. https://doi.org/10.1007/s10639-022-11357-1

[194] Tingting Wang, Shan Li, Chengyi Tan, Jianhua Zhang, and Susanne P. Lajoie. 2023. Cognitive load patterns affect temporal dynamics of self-regulated learning behaviors, metacognitive judgments, and learning achievements. *Computers & Education* 207 (Dec. 2023), 104924. https://doi.org/10.1016/j.compedu.2023.104924

[195] Patrick P. Weis and Eva Wiese. 2022. Know Your Cognitive Environment! Mental Models as Crucial Determinant of Offloading Preferences. *Human Factors* 64, 3 (May 2022), 499–513. https://doi.org/10.1177/0018720820956861 Publisher: SAGE Publications Inc.

[196] Mark Weiser. 1994. Creating the invisible interface: (invited talk). In *Proceedings of the 7th annual ACM symposium on User interface software and technology (UIST '94).* Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/192426.192428

[197] Justin D. Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward General Design Principles for Generative AI Applications. https://doi.org/10.48550/arXiv.2301.05578 arXiv:2301.05578 [cs].

[198] Justin D. Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection Not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces (IUI '21).* Association for Computing Machinery, New York, NY, USA, 402–412. https://doi.org/10.1145/3397481.3450656

[199] Justin D. Weisz, Michael Muller, Steven I. Ross, Fernando Martinez, Stephanie Houde, Mayank Agarwal, Kartik Talamadupula, and John T. Richards. 2022. Better Together? An Evaluation of AI-Supported Code Translation. In *27th International Conference on Intelligent User Interfaces (IUI '22).* Association for Computing Machinery, New York, NY, USA, 369–391. https://doi.org/10.1145/3490099.3511157

[200] Adrian Wells. 2009. *Metacognitive therapy for anxiety and depression.* Guilford Press, New York, NY, US. Pages: xvii, 316.

[201] Jennifer Wiley, Thomas D. Griffin, Allison J. Jaeger, Andrew F. Jarosz, Patrick J. Cushen, and Keith W. Thiede. 2016. Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied* 22, 4 (Dec. 2016), 393–405. https://doi.org/10.1037/xap0000096

[202] Philip H. Winne and Nancy E. Perry. 2000. Chapter 16 - Measuring Self-Regulated Learning. In *Handbook of Self-Regulation*, Monique Boekaerts, Paul R. Pintrich, and Moshe Zeidner (Eds.). Academic Press, San Diego, 531–566. https://doi.org/10.1016/B978-012109890-2/50045-7

[203] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22).* Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3491101.3519729

[204] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3491102.3517582

[205] Frank F. Xu, Bogdan Vasilescu, and Graham Neubig. 2022. In-IDE Code Generation from Natural Language: Promise and Challenges. *ACM Transactions on Software Engineering and Methodology* 31, 2 (March 2022), 29:1–29:47. https://doi.org/10.1145/3487569

[206] Nick Yeung and Christopher Summerfield. 2012. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 1594 (May 2012), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

[207] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces (IUI '22).* Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[208] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* ACM, Hamburg Germany, 1–21. https://doi.org/10.1145/3544548.3581388

[209] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3543829.3543834

[210] Lanqin Zheng, Xin Li, Xuan Zhang, and Wei Sun. 2019. The effects of group metacognitive scaffolding on group metacognitive behaviors, group performance, and cognitive load in computer-supported collaborative learning. *The Internet and Higher Education* 42 (July 2019), 13–24. https://doi.org/10.1016/j.iheduc.2019.03.002

[211] Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2022. Productivity assessment of neural code completion. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming (MAPS 2022).* Association for Computing Machinery, New York, NY, USA, 21–29. https://doi.org/10.1145/3520312.3534864

[212] Barry J. Zimmerman. 2001. Theories of self-regulated learning and academic achievement: An overview and analysis. In *Self-regulated learning and academic achievement: Theoretical perspectives, 2nd ed.* Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 1–37.

[213] Barry J. Zimmerman and Adam R. Moylan. 2009. Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education.* Routledge/Taylor & Francis Group, New York, NY, US, 299–315.

[214] Anat Zohar and Sarit Barzilai. 2013. A review of research on metacognition in science education: current and future directions. *Studies in Science Education* 49, 2 (Sept. 2013), 121–169. https://doi.org/10.1080/03057267.2013.847261 Publisher: Routledge _eprint: https://doi.org/10.1080/03057267.2013.847261.