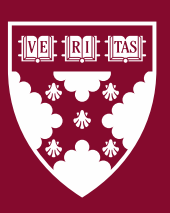
Working Paper 24-074

Don’t Expect Juniors to Teach Senior Professionals to Use Generative AI: Emerging Technology Risks and Novice AI Risk Mitigation Tactics

Katherine C. Kellogg
Hila Lifshitz
Steven Randazzo
Ethan Mollick
Fabrizio Dell'Acqua
Edward McFowland III
François Candelon
Karim Lakhani



Harvard Business School

Don’t Expect Juniors to Teach Senior Professionals to Use Generative AI: Emerging Technology Risks and Novice AI Risk Mitigation Tactics

Katherine C. Kellogg
MIT Sloan School of Management

Hila Lifshitz
Warwick Business School

Steven Randazzo
Warwick Business School

Ethan Mollick
The Wharton School

Fabrizio Dell'Acqua
Harvard Business School

Edward McFowland III
Harvard Business School

François Candelon
Boston Consulting Group

Karim Lakhani
Harvard Business School

Working Paper 24-074

Funding for this research was provided in part by Harvard Business School.

Don't Expect Juniors to Teach Senior Professionals to Use Generative AI: Emerging Technology Risks and Novice AI Risk Mitigation Tactics

Katherine C. Kellogg[1], Hila Lifshitz[2], Steven Randazzo[2], Ethan Mollick[3], Fabrizio Dell'Acqua[4], Edward McFowland III[4], François Candelon[5], and Karim Lakhani[4]

[1]MIT Sloan School of Management; [2] Warwick Business School, Artificial Intelligence Innovation Network; [3]The Wharton School, University of Pennsylvania; [4]Digital Data Design Institute, Harvard Business School; [5]Boston Consulting Group, BCG Henderson Institute

06/03/2024

Abstract

The literature on communities of practice demonstrates that a proven way for senior professionals to upskill themselves in the use of new technologies that undermine existing expertise is to learn from junior professionals. It notes that juniors may be better able than seniors to engage in real-time experimentation close to the work itself, and may be more willing to learn innovative methods that conflict with traditional identities and norms. However, this literature has not explored emerging technologies, which are seen to pose new risks to valued outcomes because of their uncertain and wide-ranging capabilities, exponential rate of change, potential for outperforming humans in a wide variety of skilled and cognitive tasks, and dependence on a vast, varied, and high volume of data and other inputs from a broad ecosystem of actors. It has also not explored obstacles to junior professionals being a source of expertise in the use of new technologies for more senior members in contexts where the juniors themselves are not technical experts, and where technology is so new and rapidly changing that the juniors have had little experience with using it. However, such contexts may be increasingly common. In our study conducted with Boston Consulting Group, a global management consulting firm, we interviewed 78 such junior consultants in July-August 2023 who had recently participated in a field experiment that gave them access to generative AI (GPT-4) for a business problem solving task. Drawing from junior professionals' in situ reflections soon after the experiment, we argue that such juniors may fail to be a source of expertise in the use of emerging technologies for more senior professionals; instead, they may recommend three kinds of *novice AI risk mitigation tactics* that: 1) are grounded in a lack of deep understanding of the emerging technology's capabilities, 2) focus on change to human routines rather than system design, and 3) focus on interventions at the project-level rather than system deployer- or ecosystem-level.

1. Introduction

In the contemporary environment of accelerating technological change, senior professionals face the dual responsibility of quickly implementing emerging technologies today (e.g., Cameron and Rahman, 2022; Karunakaran, 2024; Lebovitz, Levina, and Lifshitz, 2022; Mazmanian, Orlikowski, and Yates, 2013; Rahman and Valentine, 2021; Waardenburg, Huysman and Sergeeva, 2022), and anticipating future versions of technologies and their implications for both their clients and their own organizations (e.g., Barrett et al., 2012; Endacott & Leonardi, 2021; Karunakaran, Orlikowski and Scott, 2022; Lebovitz, Levina, and Lifshitz, 2021; Pine and Mazmanian, 2015; Sendak et al., 2020). To lead their teams and organizations in grappling with a rapidly expanding technological frontier (Dell'Acqua et al., 2023), seniors need to develop a deep understanding of new technologies and their associated capabilities (e.g., Beane and Anthony, 2023).

The literature on communities of practices tells us that, when rapid technological change and shifts in the nature of work undermine existing expertise (e.g., Beane, 2019; Bharatan, Swan and Oborn, 2022; Chan and Hedden, 2023), this poses a problem for senior members of communities of practice who have traditionally engaged in situated learning over time to become highly skilled in the use of the technologies their work requires (e.g., Barley, 1986; Brown and Duguid, 1991; Lave and Wenger, 1991; Orr, 1990). Senior professionals are often reluctant to take time away from dealing with complex problems to experiment with new technology (Anthony, 2018; Zetka, 2003). In addition, because of their senior position, they are expected to be experienced hands who know how to perform the tasks required by professionals in their role better than those more junior, so demonstrating lack of expertise during learning could threaten their image with clients and junior professionals (Beane, 2019). Finally, senior professionals understand the social rules, norms, and values through which a person's worthiness to the group is judged by members of that group better than those more junior do (e.g., Bechky, 2006; Lee, Mazmanian and Perlow, 2020), so seniors may be less willing to deviate from traditional norms to support the values consistent with new methods (Kellogg et al., 2021).

Thus, senior professionals may attempt to learn from junior professionals, who may be better able than their senior counterparts to learn to effectively use the new technologies for several reasons (Beane and Anthony, 2024). First, junior professionals are often closest to the work itself, because they are the ones engaging in concrete and less complex tasks (e.g., Bharatan, Oborn, and Swan, 2024; Karunakaran, 2018; Pine and Mazmanian, 2017; Rahman and Barley, 2017). Second, junior professionals may be more able to engage in real-time experimentation with new technologies, because they do not risk losing their mandate to lead if those around them, including clients, as well as those more junior to them, recognize that they lack the practical expertise to support their hierarchical position (e.g., Beane, 2019; Endacott & Leonardi, 2022; Leonardi, 2007). Third, junior professionals may be more willing to learn new methods that conflict with existing identities (e.g., Nelson and Irwin 2014; Lifshitz, 2018), practices (e.g., Beane and Orlikowski, 2015; Mazmanian and Beckman, 2018) and frames (e.g., Anthony and Tripsas 2016; Mazmanian, 2013).

For example, when CT scanners arrived in radiology, senior radiologists learned to effectively use them from junior technicians who were closest to the work of injecting dyes and constructing images (Barley, 1986). When help-desk queuing technology arrived in IT, senior technicians responding to help-desk calls began to ask for help from junior technicians who had engaged in experimentation with the new technology and developed skill in its use and features (Leonardi, 2007). And when patient referral and tracking technology arrived in primary care, senior medical assistants learned from those more junior how to follow each colonoscopy patient's progress through procedure and follow-up, because the juniors who were more willing to learn new methods that conflicted with existing identities, norms, and frames (Kellogg et al., 2021).

Yet, while junior professionals' teaching of the effective use of new technology to more senior professionals can be successful, its effects are far from inevitable or uniform. Thus, scholars of technology, work, and occupations have explored the question of how and when junior professionals may *fail* to be a source of expertise in the use of new technologies for more senior members. Their studies

show that status threat is often the key obstacle to junior professionals successfully coaching more senior professionals in the effective use of new technology (e.g., Beane and Anthony, 2024; Barley, 1986). Juniors coaching seniors in new technology use challenges senior members' status, which is based on the historical distinctions of the performance of higher level, more complex tasks (e.g., Leonardi & Bailey, 2008; Van Maanen and Schein, 1978), the demonstration of expertise in performing these tasks (e.g., Anthony, 2021; Bharatan, Swan & Oborn, 2022), and the enactment of traditional identities, frames, and temporal rhythms while performing the tasks (e.g., Bechky, 2003; Lifshitz, 2018; Oborn and Barrett, 2021).

This literature has been critical in highlighting that, when rapid technological change and shifts in the nature of work undermine existing expertise, senior professionals who need to rapidly upskill themselves in new technologies may turn to junior professionals. However, this literature has not explored technologies such as artificial intelligence, data analytics, social media, digital platforms, blockchain, and 3-D printing (e.g., Polykarpou, Barrett, and Oborn, 2020; Arora et al., 2023; Rahman, Karunakaran and Cameron, 2024), which scholars have labeled "emerging technologies," (e.g., Bailey et al., 2022) because of their uncertain and wide-ranging capabilities, exponential rate of change, potential for outperforming humans in a wide variety of skilled and cognitive tasks, and dependence on a vast, varied, and high volume of data and other inputs from a broad ecosystem of actors. These emerging technologies are seen to be associated not only with the risk of negative outcomes, but also with risks that are novel, unfamiliar, and involve considerable uncertainty (Arora et al., 2023; Barrett et al., 2024).

Given the rapid introduction of and change in these emerging technologies, situations in which senior professionals may need to learn to effectively use emerging technologies associated with novel and unfamiliar risks are increasingly common (Bailey et al., 2022). For example, senior professionals need to learn to mitigate risks associated with reduced control through the phases of the service encounter as they learn to effectively use platform technologies (Rahman and Valentine, 2021; Cameron & Rahman, 2022). They need to learn to mitigate risks associated with reliability breakdowns as they learn to effectively use cloud-based technologies (Karunakaran, 2021). They need to learn to mitigate risks associated with opacity as they learn to effectively use AI-based technologies (Berente et al., 2021; Lebovitz, Levina & Lifshitz, 2021; Lebovitz, Lifshitz & Levina, 2022). And they need to learn to mitigate risks associated with big data surveillance (Brayne, 2017; Rahman, Karunakaran and Cameron, 2024) and opportunistic data curators as they learn to effectively use predictive analytic technologies (Waardenburg et al., 2022). As they learn to effectively use these technologies, they are likely to turn to the juniors with whom they work for help.

This may be particularly problematic in the case of emerging technologies. The current literature on junior professionals being a source of expertise in the use of new technologies for more senior members has not explored contexts where the juniors themselves are not technical experts, and where technology is so new and rapidly changing that juniors have had no formal training on how to use the technology, have had no experience with using it in the work setting, and have had little experience with using it outside of the work setting. In such contexts, it seems unreasonable to expect that juniors should have a deep level of technical understanding of the technology.

Yet, such contexts—in which seniors need to learn to use emerging technologies associated with novel and unfamiliar risks, and in which the juniors these seniors work with may gain access to the technology even when these juniors do not have technical backgrounds and have had little training or practical experience with the technology—may be increasingly common. Emerging technologies associated with novel and unfamiliar risks are increasingly being introduced into organizations (e.g., Bailey et al., 2022). And emerging technologies such as generative AI can be easily accessed and customized without coding and without owning infrastructure, so are available to users of varying skills and technical backgrounds (e.g., Schneider et al., 2024). As we will show below, the juniors in our study expected that they would need to educate seniors about effective use of the new technology. Thus, we suggest that the research question of how and when junior professionals may fail to be a source of expertise in the use of new technologies for more senior members needs to be explored within such a context.

In our study conducted with Boston Consulting Group, a global management consulting firm, we explored this question in an experimental setting. We interviewed 78 junior consultants in July-August 2023 who had recently participated in a field experiment that gave them access to generative AI (GPT-4) for a business problem solving task. These junior professionals had been asked to solve a business problem (to identify channels and brands that would help a fictional company optimize its revenue and profitability) using fictitious interview notes with company executives and historical business performance data. Soon after they had completed the problem-solving task, we conducted semi-structured, 60-minute interviews (Spradley 1979) via Zoom with them. These juniors who had been given access to GPT-4 in the experiment were not technical experts, had had no formal training on how to use the technology, had no experience with using it in the work setting, and had little experience with using it outside of the work setting.

In a set of early background interviews with similar junior consultants, we had learned that juniors themselves were excited to use GenAI in their work, but that they anticipated that they would need to educate their managers in the use of GenAI to alleviate managers' concerns about the risks that GenAI posed to outcomes that managers valued. To better understand this phenomenon, in the 78 interviews we conducted in July-August 2023, we asked these junior consultants: 1) Can you envision your use of GenAI creating any challenges in your collaboration with managers? (For each challenge they mention, ask directly after discussing it): What are some ways to deal with these challenges? How do you think these challenges could be mitigated?

Our interviews revealed two findings that run counter to the existing literature. First, the tactics that the juniors recommended to mitigate their seniors' concerns ran counter to those recommended by experts in GenAI technology at the time, and so revealed that the junior professionals might *not* be the best source of expertise in the effective use of this emerging technology for more senior members. Second, while the literature suggests that the key obstacle to juniors educating seniors in the use of new technology is status threat to seniors, the junior consultants we interviewed reported that the key obstacle to their ability to educate their managers would be risks that GenAI posed to outcomes that managers valued (accuracy of outputs, explainability of outputs, outputs that take into account relevant contextual data, and technology users' active engagement with and interrogation of the outputs), rather than status threat associated with juniors coaching seniors.

In the sections that follow, we describe the two key theoretical arguments derived from our study. Detailed empirical data supporting these theoretical arguments is provided in **Appendix, Tables A-C**. We argue that junior professionals may fail to be a source of expertise in the effective use of an emerging technology, because they may recommend three kinds of *novice AI risk mitigation tactics* that: 1) are grounded in a *lack of deep understanding of the emerging technology's capabilities*, 2) focus on *change to human routines rather than system design*, and 3) focus on interventions at the *project-level rather than system deployer- or ecosystem-level*. We further argue that, to explain the conditions under which junior professionals may fail to be a source of expertise in the use of an emerging technology for more senior members, we must take into account the *perceived risks that the emerging technology poses* to outcomes that seniors value.

2. Bringing *emerging technology risks* into our understanding of barriers to juniors teaching seniors to effectively use new technology

As the existing literature on communities of practice would lead us to expect, under these conditions, the junior professionals we interviewed expected that they would likely be the ones to educate the senior professionals about how to effectively use the new technology (**Appendix Table A**). Yet, contrary to what the current literature would expect, juniors did not focus on status threat to seniors as the key obstacle to their ability to coach seniors in effective use of the new technology.

Instead, juniors noted that the key obstacle to their ability to coach seniors in effective use of the new technology would be the *novel risks that the technology posed* to outcomes that seniors valued (**Appendix, Table B**). Juniors highlighted that seniors would be concerned that GenAI posed a risk to the

accuracy of outputs (the degree to which outputs have attributes that correctly reflect the true value of the intended attributes of a concept or event in a particular context of use; Steimers & Schneider, 2022); to the *explainability of outputs* (the degree to which the outputs are presented in a way that is understandable for humans; Steimers & Schneider, 2022); and to the *contextualization of outputs* (the degree to which outputs are coherent, relevant, or in compliance with particular constraints or rules due to absorbing and taking into account pertinent contextual data when producing outputs; Sai et al., 2024). Juniors also expected that seniors would be concerned that GenAI posed a risk to user engagement by promoting *automation complacency* (in which users may trust the outputs provided by generative AI in situations where they should not; Van Dis et al., 2023).

We turned to the literature on the risks associated with generative AI technology to better understand the kinds of risks that GenAI was expected to pose to seniors' valued outcomes of accuracy of outputs, explainability of outputs, outputs that took into account relevant contextual data, and GenAI users' active engagement with and interrogation of the outputs. The literature suggested that there were several unique characteristics of this emerging technology that posed risks to these valued outcomes (**Paper Table 1**) that seemed to be different from the technologies that have been examined by scholars in the existing community of practice literature on juniors coaching seniors.

First, *GenAI can be accessed and customized by novice users without coding and without owning infrastructure*. Prior studies of juniors teaching seniors have analyzed technologies in which interacting with the technology required infrastructure and raised barriers for ordinary users. For example, the CT scanning technology that Barley (1986) studied required a large investment in hardware by hospital administrators, and integration with existing hospital medical equipment and IT infrastructure by IT professionals. In contrast, GenAI allows novice users access the technology directly from their computers and collaborate with it in a nearly instantaneous fashion.

Second, *GenAI has uncertain and wide-ranging capabilities and is changing at an exponential rate* (OpenAI, 2023a; Webb et al., 2023; Wei et al., 2022). Prior studies of juniors teaching seniors have examined technologies purpose-built for a specific application, and changing at a slower rate. For example, the Davinci robot that Beane (2019) studied was purpose-built to allow a single surgeon to perform both support (retraction) and direct surgical action (dissection) to an unprecedented degree. In contrast, with GenAI, a broad range of applications can be performed by a single system, and GenAI's capabilities are expanding exponentially.

Third, *GenAI carries the possibility of outperforming humans in a wide variety of skilled and cognitive tasks* (Bubeck et al., 2023). Prior studies of juniors teaching seniors have investigated technologies designed to assist professionals. For example, the EMR technology that Kellogg and colleagues (2021) studied was able to assist medical assistants to do their work of patient check-in and check-out processes and referrals and insurance authorizations more accurately and quickly. In contrast, GenAI holds the potential to surpass human performance across a diverse range of skilled and cognitive activities.

Fourth, *GenAI combines data and other inputs at an unprecedented scale and detail from a broad ecosystem of actors*. Prior studies of juniors teaching seniors have analyzed technologies that involved model developers, system deployers, and system users, but were dependent on a limited set of data from each. For example, the Factset and CapIQ technology that Anthony (2021) studied gathered data directly from public SEC filings, automatically calculated particular key metrics, and then fed these calculations directly into spreadsheets in the user's organization. In contrast, with GenAI, continuous change in data sources at an unprecedented scale and detail affect what the technology is able to do. Thus, GenAI emerges through a set of expanded relations and continues to emerge in new ways as those relations evolve.

In sum, one of our findings that runs counter to the existing literature is that juniors expected that the key obstacle to their ability to coach seniors in effective use of GenAI would be the *novel risks that this emerging technology posed* to outcomes that seniors valued. Thus, to explain how and when junior professionals may fail to be a source of expertise in the use of an emerging technology for more senior members, we must take into account not only status threat, but also risks to valued outcomes that are

posed by emerging technologies that *1) have uncertain and wide-ranging capabilities and are changing at an exponential rate, 2) have the possibility of outperforming humans in a wide variety of skilled and cognitive tasks,* and *3) combine data and other inputs at an unprecedented scale and detail from a broad ecosystem of actors.* Such risks associated with an emerging technology may pose threats to seniors' valued outcomes of accuracy of outputs, explainability of outputs, outputs that take into account relevant contextual data, and technology users' active engagement with and interrogation of the outputs.

3. Bringing *novice AI risk mitigation tactics* into our understanding of juniors teaching seniors to effectively use new technology

A second finding that runs counter to the existing literature is that the tactics that the juniors recommended to mitigate the risks that GenAI posed to seniors' valued outcomes ran counter to the risk mitigation tactics recommended by experts in GenAI technology at the time, and so revealed that the junior professionals might *not* be the best source of expertise in the effective use of the emerging technology for more senior members.

We found both the human-computer interaction (HCI) literature's concept of *novice tactics for technology use* and recent publications by experts in GenAI to be helpful to understanding the dynamics we observed. The HCI literature on novice tactics for technology use helps to explain why junior professionals' time-tested ways of self-upskilling by engaging in real-time experimentation may not allow them to identify the most productive ways to mitigate the risk threats posed by an emerging technology. Theorists of human-computer interaction highlight that novice programmers often intuitively approach interactions with a new technology in counterproductive ways (e.g., Ko et al., 2004; Lahtinen et al., 2005). HCI researchers are interested in surfacing pitfalls that non-experts are susceptible to, in order to inform the design of technologies that are easier to effectively use by people who are not formally trained in their use (e.g., Yang et al., 2018; Zamfirescu-Pereira et al., 2023).

We compared juniors' suggested AI risk mitigation tactics to those recommended by GenAI experts in the emerging literature at the time. We found that the juniors we studied recommended tactics for AI risk mitigation that differed in three key ways from tactics recommended by experts at the time. These *novice AI risk mitigation tactics* included: 1) *tactics that stemmed from a lack of deep understanding of the characteristics of the emerging technology*, 2) tactics informed by juniors' past experience with human-to-human interaction that *overestimated the potential of changing human routines (rather than system design)* to mitigate these risks, 3) tactics informed by juniors' past experience with project-level work that *overestimated the potential of making changes at the project level (rather than the deployer- or ecosystem-level)* to mitigate these risks. We elaborate each of these tactics below. The tactics are summarized in **Paper Table 2**, and examples of juniors' descriptions of each of the tactics are presented in **Appendix Tables C1-C3**.

3.1 Novice tactics type #1: Lacking deep understanding the characteristics of the emerging technology (Appendix Table C1)

The first type of novice tactics we discovered was juniors' recommendations for mitigating GenAI output risks related to accuracy, explainability, and contextualization which demonstrated juniors' lack of a deep understanding of the capabilities of GenAI technology. Juniors' lack of deep understanding is not surprising since, at the time of the study, these juniors were novice users. They were not technical experts, and the technology was so new and rapidly changing that juniors had had no formal training and little practical experience with using it.

Yet, it is important to understand this tactic, because the current literature does not examine contexts in which seniors need to learn to use emerging technologies that can be accessed and customized by novice users without coding and without owning infrastructure; and such contexts are increasingly common. Juniors in our study expected that they would need to educate seniors about effective use of the new technology (**Appendix, Table A**).

3.1.1. Some juniors lacked a deep understanding of GenAI accuracy. Some consultants did not understand that the accuracy of GenAI was ultimately limited at the time of the study. Thus, these juniors recommended mitigating the risk of GenAI output inaccuracy by "Using a standardized way of asking questions," or "Doing the work yourself first. Only use Gen AI once you have any kind of preliminary output, and augment only versus create."

In contrast, papers by GenAI experts available at the time demonstrated that generative AI models can confidently present users with information that is hallucinatory, and that this can result in incorrect output that does not accurately reflect real people, places, or facts (e.g., Weidinger et al., 2022). Hallucinations can occur when the model tries to fill in gaps in its knowledge or when the input is ambiguous. Thus, experts recommended addressing GenAI's accuracy issues by deciding on appropriate use cases where error risks are acceptable (e.g., Open AI, 2023a), and by independently testing GenAI's reliability in executing each subtask; for example by breaking down the users' needed subtasks (e.g., "information gathering") and creating evaluations for each independently (OpenAI, 2023b).

3.1.2. Some juniors lacked a deep understanding of GenAI explainability. Some consultants did not understand that the explainability of GenAI was not possible at the time of the study. Thus, these juniors recommended tactics for "explaining its rationale" to managers, and "understanding the source of the recommendation or the result, being able to explain it," so that managers could better understand it.

In contrast, papers by GenAI experts available at the time demonstrated that LLMs do not provide transparency for their reasoning (e.g., Jain, Sarthak and Byron, 2019; Jacovi and Goldberg, 2020). LLM models are complex and opaque, because the deep neural networks that underpin them are composed of billions of parameters, leading to emergent behaviors that are often unpredictable and not easily interpretable (Lin et al., 2023). Experts further showed that GenAI provides an illusion of transparency, but that what GPT gives for explanation does not always match what models actually focus on to provide their outputs (e.g., Jain, Sarthak and Byron, 2019; Dasgupta et al., 2022). Indeed, sometimes models do not actually rely on the chains-of-thought they purport to when reasoning, so relying on these may create a false sense of security in the user (Turpin et al., 2023). Thus, experts recommended avoiding GenAI use where a high degree of explainability is required (Bender et al., 2021; Liu et al., 2023). They also recommended providing the user with global explanations about the model logic and about how to improve the input, since it is not possible to explain the model process for a specific output (Liao and Vaughn, 2023).

3.1.3. Some juniors lacked a deep understanding of GenAI contextualization. Finally, some consultants believed that GenAI was not capable of contextualization. Thus, these juniors recommended tactics such as "only using GenAI for cases where contextualization is not necessary."

In contrast, papers by GenAI experts available at the time showed that GenAI is strong at contextualization, with the appropriate prompting methods. They noted that effective communication with generative AI requires providing contextual information, and specifying the desired output (e.g., Zhou et al., 2023). They also recommended using Retrieval-Augmented Generation (RAG) methods to complement the general knowledge of LLMs with internal data to provide contextually relevant answers (e.g., Lewis et al., 2020; Gu et al., 2020). For example, a consultant could ask a RAG AI agent to provide a summary of internal documents related to a customer organization.

3.2 Novice tactics type #2: Overestimating the potential of changing human routines (rather than system design; Appendix Table C2)

The second type of novice tactics we discovered was juniors overestimating the potential for mitigating GenAI output risks related to accuracy, explainability, and contextualization by changing human routines (rather than system design).

3.2.1 Juniors recommended change to human routines rather than system design to mitigate accuracy risks. Some consultants suggested mitigating output risks related to accuracy by "training users to validate results" and by having managers "review our prompts and responses."

In contrast, papers by GenAI experts available at the time showed that expecting humans to validate user prompts is very difficult to do, because human users may not always have time to go through activity logs at the speed or scale they desire, or may "fall asleep at the wheel"—fail to exert effort and remain attentive, allowing the AI to substitute, rather than augment their performance (Dell'Acqua, 2022).

For this reason, experts suggested mitigating output risks related to accuracy by making changes to system design. This could be done by fine-tuning a model's parameters based on additional, specialized data (e.g., Devlin et al., 2019; Lee et al., 2020). It could also be done by setting up automatic monitoring with a second system (such as a classifier, or a generative AI system capable of producing its own chains-of-thought; Saunders et al., 2022) And it could be done by using models that provide links to sources or using more accurate models that improve accuracy by combining retrieved data with the generative content from LLMs, with meticulous claim-by-claim fact checking (Min et al., 2023).

3.2.2 Juniors recommended change to human routines rather than system design to mitigate complacency risks. Some consultants overestimated the potential of changing human routines, rather than system design, to address the risk of complacency, for example, by suggesting teaching consultants to "know that ownership is theirs," and teaching managers to "push back on consultants who seem to be just copying and pasting from GPT."

While GenAI experts would not disagree with these methods, expert recommendations at the time also showed that juniors' focus on changing human routines ignored system design-focused approaches effective for mitigating complacency risk. For example, experts noted that a GenAI system could be designed to provide proactive self-reflective prompts to help end-users calibrate their confidence in system outputs, asking, "How confident are you in understanding this output? Does anything require explanation?" (Gmeiner et al., 2023). A GenAI system could also be designed with an interface that visualizes uncertainty, because highlighting uncertain content can build awareness that AI-generated content may be wrong (Vasconcelos et al., 2023b). Further, a GenAI system could support pattern-matching between GenAI suggestions and users' task goals; for example, the system's output could have keywords highlighted, such as variable names or function calls, that would indicate code fit (Barke et al., 2023). Finally, the models' default behavior could apply hard-coded restrictions to improve an LLM's alignment to a custom objective (Lu et al., 2022).

3.2.3. Juniors recommended change to human routines rather than system design to mitigate contextualization risks. Some consultants focused on mitigating contextualization risks by "gaining agreement within the team around using GenAI for documents that did not require a high degree of contextualization." Other consultants recommended mitigating contextualization risks by "training consultants in prompt engineering."

While GenAI experts agreed that training users in prompt engineering is an accepted method for improving contextualization, experts also highlighted system-based approaches for improving contextualization. In addition to using RAG, as described above, experts recommended aligning the model's generations to particular objectives specified by users by using human-labeled preference data at the fine-tuning stage (e.g., Song et al., 2023), designing GenAI systems to begin with a prompt to the user to communicate their goals and preferences to the system (Zamfirescu-Pereira et al., 2023) and creating prompts centrally and building these into the system (e.g., Achiam et al., 2023).

3.3. Novice tactics type #3: Overestimating the potential of intervening at the project level (rather than system deployer- or ecosystem-level; Appendix Table C3)

The third type of novice tactics we discovered was juniors' overestimating the potential for intervening at the project level (rather than at the system deployer or ecosystem level) to mitigate GenAI output risks related to accuracy, explainability, and contextualization. It is not surprising that juniors who were working at the project level with their managers thought about mitigating challenges at that level, rather than at the system deployer- or ecosystem-level. Yet, it is important to understand this tactic, because the current literature does not highlight how this may mean that juniors may not be the best source of expertise regarding mitigating risks related to the use of an emerging technology.

3.3.1 Juniors recommended interventions at the project level rather than deployer- or ecosystem-level to mitigate accuracy risks. Some consultants suggested mitigating output risks related to accuracy at the project level by having "consultants and managers agree on the conditions under which GenAI can be used reliably" or by "having managers review consultants' work process" around working with GenAI.

In contrast, papers by GenAI experts available at the time highlighted that, because LLM developers are often not system deployers, and because LLMs provide increased accessibility of powerful models by users of varying skills and technical backgrounds, it is also important to intervene at the system deployer- and LLM developer-level to mitigate accuracy risk.. Experts recommended several actions that could be taken by system deployers to mitigate accuracy risks: a) communicate to users the intended conditions under which GenAI can be used reliably (Liao and Vaughn, 2023); b) provide co-audit tools (Mündler et al. 2023); c) create a prompt library (Svendsen et al., 2023); d) continually assess the alignment of LLMs vis-à-vis evaluation metrics (e.g., Otani et al., 2023); and e) develop evaluation methods that take into account emergent capabilities as models get more capable, and are open ended enough to detect unforeseen risks (OpenAI, 2023a). For example, system deployers could provide users with access to ChatProtect, an AI-based co-audit tool with features to detect and remove hallucinated content from generated text. The co-audit experience lets the user inspect different sentences to detect hallucinations via sampling multiple times from the LLM (Mündler et al. 2023). Another approach recommended by experts was for system deployers to develop robust evaluation frameworks, and create a monitoring role within the organization to continually assess the alignment of LLMs vis-à-vis these evaluation metrics (e.g., Otani et al., 2023).

Experts also recommended intervening at the ecosystem level to mitigate accuracy risks. Experts highlighted the critical dependency of LLMs on their training data, which, while comprehensive, often fails to encapsulate the rapid changes inherent in real-world contexts, particularly in scenarios post-dating the training period. Thus, experts suggested that LLM developers should assess the representativeness, robustness, and quality of their data sources, and implement mechanisms that allow LLMs to continually learn from new data, in order to capture recent developments and trends (Li et al., 2023). Experts suggested that developers could do this by creating methods such as dynamically adjusting LLM behaviors by implementing algorithms to assess the recency and applicability of data points (Meng et al., 2021).

Experts also suggested that LLM developers could help to mitigate accuracy risks by: being clear and upfront about how well the GenAI system performs different tasks by explaining its capabilities and limitations (OpenAI, 2023c; Solaiman et al., 2023), by identifying the source materials used to generate it (Liao and Vaughan, 2023), and by flagging and correcting misleading outputs (Ahmad et al., 2023). Further, experts suggested that model developers should refrain from anthropomorphizing the AI, because this may lead people to falsely assume that the AI has a greater overall accuracy than it actually does (Vasconcelos et al., 2023a).

3.3.2 Juniors recommended interventions at the project level rather than deployer- or ecosystem-level to mitigate complacency risks. Some consultants suggested mitigating output risks related to

complacency at the project level by managers "giving extra time for quality checking," and "not short selling projects" based on expected time savings.

While experts would not disagree with these methods for mitigating complacency risk, they also highlighted actions that can be taken at the system deployer- and ecosystem-level to mitigate output risks related to complacency. At the system deployer-level, to address this risk, LLM deployers can onboard users to the system (De-Arteaga et al. 2020); they can also provide personalized adjustments for users by assessing users' confidence in their own abilities and adjusting the user experience to help overconfident users develop appropriate reliance (Lu & Yin 2021). Further, system deployers can design interfaces with cognitive forcing functions that reduce complacency such as time-outs, on-demand explanations, and asking users to explicitly rule out alternatives (Buçinca et al. 2021). Finally, system deployers can allow users to clearly specify how tasks are allocated between the human and system, to allow users to better distribute the workload according to the respective strengths and weaknesses of humans and GenAI, and to reduce the cognitive demand on users trying to discern the relative responsibilities on a moment-by-moment basis (Simkute et al., 2023).

Experts also recommended interventions at the ecosystem level to mitigate complacency risks. They noted that GenAI developers can mitigate complacency by building their models to be stringent in rejecting requests that go against their content policy, and by considering the impact on users of the model's style, tone, or perceived personality (OpenAI, 2023a). Further, they suggested that system feedback could highlight prompt changes and the resulting output changes (Zamfirescu-Pereira et al., 2023). Finally, they noted that developers could support debugging, for example, by providing test cases and test data that users could employ to identify corner cases (Vaithilingam et al., 2022).

4. Discussion

Prior literature on technology and learning assumes that junior professionals are better positioned to learn how to effectively use new technologies than are seniors, and that the key barrier to junior professionals being a source of expertise for seniors is the status threat associated with juniors coaching seniors. Our analysis suggests otherwise. Here, we expand our discussion of the concepts of *novice AI risk mitigation tactics* and *emerging technology risk threats* to explore how these concepts can inform our understanding of juniors and seniors learning to use emerging technologies.

4.1 Novice AI risk mitigation tactics

We show that juniors, rather than being a source of expertise for senior professionals in the effective use of emerging technologies, may instead recommend three kinds of *novice AI risk mitigation tactics* for addressing risks to valued outcomes that: 1) are grounded in a lack of deep understanding of the emerging technology's capabilities, 2) focus on change to human routines rather than system design, and 3) focus on interventions at the project-level rather than system deployer- or ecosystem-level.

Juniors may recommend novice AI risk mitigation tactics because juniors themselves may not be technical experts, and because when technology is nascent and exponentially changing, juniors may have had no formal training on how to use the technology, no experience with using it in the work setting, and little experience with using it outside of the work setting.

Because emerging technologies have uncertain and wide-ranging capabilities that are changing at an exponential rate, juniors may not be fully informed about their capabilities. Because emerging technologies have the potential for outperforming humans in a wide variety of skilled and cognitive tasks, juniors' focus on change to human routines may be less effective in mitigating risks than would be a focus changes to system design. And, because emerging technologies depend on a vast, varied, and high volume of data and other inputs from a broad ecosystem of actors, juniors' focus on interventions at the project-level may be less likely to be effective than interventions at the system deployer- and ecosystem-level.

4.2 Emerging technology risk threats

We further show that status threat may not be the primary reason that juniors are thwarted in their attempts to coach more senior members in the use of emerging technologies. We do not disagree that junior professionals coaching seniors in the use of new technology may challenge seniors' status, which is based on the historical distinctions of the performance of higher level, more complex tasks, the demonstration of expertise in performing these tasks, and the enactment of traditional identities and frames while performing the tasks. Instead, we argue that, in the case of learning to use emerging technologies, seniors may be more concerned about the risks that emerging technologies pose to their valued outcomes than to the risks that juniors coaching them in technology use poses to seniors' status preservation.

Emerging technologies pose new risks when they 1) have uncertain and wide-ranging capabilities and are changing at an exponential rate, 2) have the potential to surpass human performance in various skilled and cognitive tasks, and 3) combine data and other inputs at an unprecedented scale and detail from a broad ecosystem of actors. Such risks may bring threats to seniors' valued outcomes of accuracy of outputs, explainability of outputs, outputs that take into account relevant contextual data, and technology users' active engagement with and interrogation of the outputs.

4.3 Implications for research on professionals learning to use emerging technologies

For research on professionals learning to use emerging technologies, the inclusion of the concept of *novice AI risk mitigation tactics* can enhance our understanding of how juniors may fail to be a source of expertise for seniors in the effective use of emerging technologies because of 1) a lack of deep understanding of the emerging technology's capabilities, 2) past experience with human-to-human interaction that leads them to focus on change to human routines rather than system design, and 3) past experience at the project level that leads them to focus on interventions at the project level rather than system deployer- or ecosystem- level.

Our study has several limitations that could be addressed in future studies of professionals learning to use emerging technologies. First, we conducted an experiment with a relatively small set of juniors drawn from one consulting firm, and this set of professionals may not be representative of other kinds of professionals. For example, while the current literature emphasizes status threat as a barrier to juniors serving as a source of expertise for seniors, this threat may be unlikely in a consulting firm where junior worker performance is not a threat to the career path for their seniors due to highly structured, up-or-out, two-year promotion cycles; in this context, senior's career prospects are not impacted by the performance of their juniors. Second, the time we gave participants in the experiment to engage with GenAI was very limited; it is quite likely that, once given more exposure and training, these same juniors would develop more effective tactics for GenAI use. Third, at the time of the experiment, the technology was still nascent; consultants were in a first exposure moment, so the speed at which the technology was evolving was likely and understandably not apparent to them; juniors might respond differently if they knew that the technology was exponentially changing. Future research could examine a larger population of professionals, from a wider variety of organizations and industries, after they have spent more time engaging with the technology, and after they have realized the degree to which the technology is exponentially changing. Such research could examine which of the challenges we identified persist, and what new challenges emerge. Fourth, the emerging technology we studied was generative AI. As noted, this technology can be easily accessed and customized without coding and without owning infrastructure. Future research could examine the ways in which the challenges and mitigation challenges we elaborate are similar and different for other kinds of emerging technologies that are not as easily accessed and customized.

4.4 Implications for research on novice tactics for technology use

Theorists of human-computer interaction highlight that novices often bring default behaviors, intuitions, preferences, and capabilities to technology use that lead them to approach interactions with technology in counterproductive ways (e.g., Lahtinen et al., 2005). These theorists show, for example,

that novice end users may bring human intuitions rooted in social experiences to technology use that lead these novices to attempt debugging opportunistically rather than systematically (e.g., Ko et al., 2004) or to attempt to design an ML model by directly mapping a personal need to a model task rather than by framing an achievable task (e.g., Yang et al., 2018). This literature suggests that designers can build future tools that better support users in the face of these common counterproductive intuitions and struggles.

In some ways, our research has similar implications. We point to the importance of end-user focused tool design interventions such as "Design a system that provides the end user with proactive self-reflective prompts," and "Build an interface that visualizes uncertainty for the end user."

However, our concept of *emerging technology risk threats* also suggests that novice end users may engage with GenAI technology in less productive ways not only because of their own novice intuitions, but also because it is extremely difficult to mitigate risks associated with emerging technologies that have uncertain and wide-ranging capabilities that are changing at an exponential rate, that carry the possibility of outperforming humans in a wide variety of skilled and cognitive tasks, and that combine data and other inputs at an unprecedented scale and detail from a broad ecosystem of actors. We contend that the HCI literature's "design-to-help-novices" approaches, while important, may not be sufficient in addressing the full consequences of these emerging technology risk threats.

Thus, we argue that studies need to move beyond a focus on the design of human-computer interaction, and focus also on the system deployer- and ecosystem-levels in which the technologies are developed, deployed, and used. This would include research on the implications of the system-deployer level interventions, such as setting policies that outline appropriate use cases where error risks are acceptable, and continually assessing the alignment of LLMs vis-à-vis robust evaluation metrics. It would also include research on the implications of the ecosystem-level interventions, such as encouraging developers to adjust LLM behaviors by implementing algorithms to assess the recency and applicability of data points, and to avoid anthropomorphizing the AI. By taking such a broader perspective, we may be better able to identify interventions related to a larger range of human actors (system developers, system deployers, and system users) and material actors (data, models, and infrastructure) that could be more effective for mitigating risks to humans' valued outcomes.

4.5 Implications for practice

As corporate leaders attempt to keep their employees' skills in sync with emerging technologies such as artificial intelligence, blockchain, and 3-D printing, they may imagine that juniors will be able to help seniors to learn to use new technology. In contrast, we highlight that learning to use emerging technologies requires addressing a novel set of risks, and that juniors themselves may be novices in technology use when technology is nascent and exponentially changing; thus, corporate leaders should make sure that both their junior and senior members are focused on addressing AI risks by deeply understanding the emerging technology's unique capabilities and limitations, making changes to system design in addition to human routines, and intervening in sociotechnical ecosystems, rather than only at the project level.

5. REFERENCES

Anthony, C. (2018). To question or accept? How status differences influence responses to new epistemic technologies in knowledge work. *Academy of Management Review*, 43(4), 661-679.

Anthony, C. (2021). When knowledge work and analytical technologies collide: The practices and consequences of black boxing algorithmic technologies. *Administrative Science Quarterly*, 66(4), 1173-1212.

Anthony, C., & Tripsas, M. (2016). Organizational identity and innovation. *The Oxford handbook of organizational identity*, *1*, 417-435.

Arora, A., Barrett, M., Lee, E., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization*, 33(3), 100478.

Bailey, D. E., Faraj, S., Hinds, P. J., Leonardi, P. M., & von Krogh, G. (2022). We are all theorists of technology now: A relational perspective on emerging technology and organizing. *Organization Science*, *33*(1), 1-18.

Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly*, 78-108.

Barke, S., James, M. B., & Polikarpova, N. (2023). Grounded copilot: How programmers interact with code-generating models. Proceedings of the ACM on Programming Languages, 7(OOPSLA1), 85-111.

Barrett, M., Oborn, E., Orlikowski, W. J., & Yates, J. (2012). Reconfiguring boundary relations: Robotic innovations in pharmacy work. *Organization Science*, 23(5), 1448-1466.

Barrett, M., Oborn, E., Prince, K., Lee, E. (2024). A relational perspective on digital technologies and organizing risk: Telemedicine as a risk object in the re-imagining of healthcare futures. Working paper.

Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, 64(1), 87-123.

Beane, M., & Anthony, C. (2024). Inverted apprenticeship: How senior occupational members develop practical expertise and preserve their position when new technologies arrive. *Organization Science*, 35(2), 405-431.

Beane, M., & Orlikowski, W. J. (2015). What difference does a robot make? The material enactment of distributed coordination. *Organization Science*, 26(6), 1553-1573.

Bechky, B. A. (2003). Sharing meaning across occupational communities: The transformation of understanding on a production floor. *Organization Science*, *14*(3), 312-330.

Bechky, B. A. (2006). Gaffers, gofers, and grips: Role-based coordination in temporary organizations. *Organization Science*, *17*(1), 3-21.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3).

Bharatan, I., Swan, J., & Oborn, E. (2022). Navigating turbulent waters: Crafting learning trajectories in a changing work context. *Human Relations*, 75(6), 1084-1112.

Bharatan, I., Oborn, E., & Swan, J. (2024). Finding your Sea Legs: Exploring Newcomer Embodied Learning in an Extreme Context. *Journal of Management Studies*.

Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review*, *82*(5), 977-1008.

Brown, J. S., & Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science*, *2*(1), 40-57.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 188:1-188:21.

Cameron, L. D., & Rahman, H. (2022). Expanding the locus of resistance: Understanding the co-constitution of control and resistance in the gig economy. *Organization Science*, 33(1), 38-58.

Chan, C. K., & Hedden, L. N. (2023). The role of discernment and modulation in enacting occupational values: How career advising professionals navigate tensions with clients. *Academy of Management Journal*, *66*(1), 276-305.

Chiang, C.-W., & Yin, M. (2021). You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. 13th ACM Web Science Conference 2021, 120–129

Dan J., and Hashimoto, T. (2023) Version 1 of "Navigating the grey area: How expressions of uncertainty and overconfidence affect language models." *arXiv preprint arXiv:2302.13439* (2023).

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*. De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24-013).

Dencik, J., Goehring, B., & Marshall, A. (2023). Managing the emerging role of generative AI in next-generation business. *Strategy & Leadership*, 51(6), 30-36.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Endacott, C. G., & Leonardi, P. M. (2022). Artificial intelligence and impression management: Consequences of autonomous conversational agents communicating on one's behalf. *Human Communication Research*, 48(3), 462-490.

Gmeiner, F., Yang, H., Yao, L., Holstein, K., & Martelaro, N. (2023, April). 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 (pp. 1-20).

Guu, Kelvin, et al. "Retrieval augmented language model pre-training." *International conference on machine learning*. PMLR, 2020

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. *arXiv preprint arXiv:2004.03685*.

Jain, Sarthak, and Byron C. Wallace. "Attention is not explanation." *arXiv preprint arXiv:1902.10186* (2019)].

Karunakaran, A. (2018). Truce structures: Examining cross-professional coordination in the wake of technological and institutional change (Doctoral dissertation, Massachusetts Institute of Technology).

Karunakaran, A. (2021). In cloud we trust? Co-opting occupational gatekeepers to produce normalized trust in platform-mediated interorganizational relationships. *Organization Science*, 33(3), 1188-1211.

Karunakaran, A. (2024). Frontline professionals in the wake of social media scrutiny: Examining the processes of obscured accountability. *Administrative Science Quarterly*, Articles in Advance.

Karunakaran, A., Orlikowski, W. J., & Scott, S. V. (2022). Crowd-based accountability: Examining how social media commentary reconfigures organizational accountability. *Organization Science*, 33(1), 170-193.

Kellogg, K. C., Myers, J. E., Gainer, L., & Singer, S. J. (2021). Moving violations: Pairing an illegitimate learning hierarchy with trainee status mobility for acquiring new skills when traditional expertise erodes. *Organization Science*, *32*(1), 181-209.

Ko, A.J. et al. 2004. Six Learning Barriers in End-User Programming Systems. Proceedings of the 2004 IEEE Symposium on Visual Languages - Human Centric Computing (USA, 2004), 199–206.

Lahtinen, E., Ala-Mutka, K., & Järvinen, H. M. (2005). A study of the difficulties of novice programmers. ACM SIGCSE Bulletin, 37(3), 14-18.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Lebovitz, S., Lifshitz, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1), 126-148.

Lebovitz, S., Levina, N., & Lifshitz, H. (2021). Is AI ground truth really true? the dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly*, 45(3).

Lee, C., Cho, K., & Kang, W. (2019). Mixout: Effective regularization to finetune large-scale pretrained language models. arXiv preprint arXiv:1909.11299.

Lee, M. Y., Mazmanian, M., & Perlow, L. (2020). Fostering positive relational dynamics: The power of spaces and interaction scripts. *Academy of Management Journal*, *63*(1), 96-123.

Leonardi, P. M. (2007). Activating the informational capabilities of information technology for organizational change. *Organization Science*, *18*(5), 813-831.

Leonardi, P. M., & Bailey, D. E. (2008). Transformational technologies and the creation of new work practices: Making implicit knowledge explicit in task-based offshoring. *MIS Quarterly*, 411-436.

Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.][

Li, J., Tang, T., Zhao, W. X., Wang, J., Nie, J. Y., & Wen, J. R. (2023). The Web Can Be Your Oyster for Improving Large Language Models. arXiv preprint arXiv:2305.10998.

Liao, Q. V., & Vaughan, J. W. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*.

Lifshitz, H. (2018). Dismantling knowledge boundaries at NASA: The critical role of professional identity in open innovation. *Administrative Science Quarterly*, *63*(4), 746-782.

Lin, Z., Trivedi, S., & Sun, J. (2023). Generating with confidence: Uncertainty quantification for black-box large language models. arXiv preprint arXiv:2305.19187.

Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Cheng, R. G. H., Klochkov, Y., ... & Li, H. (2023). Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374*.

Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., ... & Choi, Y. (2021). Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. arXiv preprint arXiv:2112.08726.

Lu, Z., & Yin, M. (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 (pp. 1– 16). Association for Computing Machinery.

Mazmanian, M. (2013). Avoiding the trap of constant connectivity: When congruent frames allow for heterogeneous practices. *Academy of Management Journal*, 56(5), 1225-1250.

Mazmanian, M., Orlikowski, W. J., & Yates, J. (2013). The autonomy paradox: The implications of mobile email devices for knowledge professionals. *Organization Science*, 24(5), 1337-1357.

Mazmanian, M., & Beckman, C. M. (2018). "Making" your numbers: Engendering organizational control through a ritual of quantification. *Organization Science*, 29(3), 357-379.

Meng, Lingheng, Rob Gorbet, and Dana Kulić. "Memory-based deep reinforcement learning for pomdps." *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021

Mündler, N., He, J., Jenko, S., & Vechev, M. (2023). Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. arXiv preprint arXiv:2305.15852.

Nelson, A. J., & Irwin, J. (2014). "Defining what we do—all over again": Occupational identity, technological change, and the librarian/Internet-search relationship. *Academy of Management Journal*, 57(3), 892-928.

Oborn, E., & Barrett, M. (2021). Marching to different drum beats: A temporal perspective on coordinating occupational work. *Organization Science*, 32(2), 376-406.

Open AI. (2023a) GPT-4 technical report. arXiv. Preprint posted online March 27, 2023. doi:10. 48550/arXiv.2303.08774

Open AI. (2023b). "Evaluating AI Agents-thoughts on this flow? Community-OpenAI Developer Forum," 2023. https://community.openai.com/t/evaluating-ai-agents-thoughts-on-this-flow/314663/1.

OpenAI, (2023c). "Gpt-4 system card," 03 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Orr, J. E. (1990). Talking about machines: An ethnography of a modern job. Cornell University.

Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., ... & Satoh, S. I. (2023). Toward verifiable and reproducible human evaluation for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14277-14286).

Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, *32*(1), 18-41.

Pine, K. H., & Mazmanian, M. (2017). Artful and contorted coordinating: The ramifications of imposing formal logics of task jurisdiction on situated practice. *Academy of Management Journal*, *60*(2), 720-742.

Pine, K., & Mazmanian, M. (2015). Emerging insights on building infrastructure for data-driven transparency and accountability of organizations. iConference 2015 Proceedings.

Polykarpou, S., Barrett, M., & Oborn, E. (2020). Place and Organizing for Emerging Technologies-Challenges of Scaling 3DPrinting Across a UK Hospital. In Academy of Management Proceedings (Vol. 2020, No. 1, p. 20450). Briarcliff Manor, NY 10510: Academy of Management.

Puthumanaillam, G., Liu, X., Mehr, N., & Ornik, M. (2023). Weathering ongoing uncertainty: learning and planning in a time-varying partially observable environment. arXiv preprint arXiv:2312.03263.

Rahman, H. A., & Barley, S. R. (2017). Situated redesign in creative occupations–An ethnography of architects. *Academy of Management Discoveries*, 3(4), 404-424.

Rahman, H. A., & Valentine, M. A. (2021). How managers maintain control through collaborative repair: Evidence from platform-mediated "gigs." *Organization Science*, 32(5), 1300-1326.

Rahman, H. A., Karunakaran, A., & Cameron, L. D. (2024). Taming platform power: Taking accountability into account in the management of platforms. *Academy of Management Annals*, 18(1), 251-294.

Sai, S., Gaur, A., Sai, R., Chamola, V., Guizani, M., & Rodrigues, J. J. (2024). Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies and Limitations. IEEE Access.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Schneider, J., Abraham, R., & Meske, C. (2024). Governance of Generative Artificial Intelligence for Companies. *arXiv preprint* arXiv:2403.08802.

Semnani, S., Yao, V., Zhang, H., & Lam, M. (2023, December). WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 2387-2413).

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., ... & O'Brien, C. (2020, January). "The human body is a black box" supporting clinical decision-making with deep learning. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 99-109).

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., ... & Robinson, D. G. (2023). Practices for Governing Agentic AI Systems. Research Paper, OpenAI, December.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180.

Solaiman, I. (2023, June). The gradient of generative AI release: Methods and considerations. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (pp. 111-122).

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., & Wang, H. (2023). Preference Ranking Optimization for Human Alignment. *arXiv e-prints*, arXiv-2306.

Spradley, J. P. (1979). Interviewing an informant. *The ethnographic interview*, 55-68.

Steimers, A., & Schneider, M. (2022). Sources of risk of AI systems. *International Journal of Environmental Research and Public Health*, 19(6), 3641.

Svendsen, A., & Garvey, B. (2023). An Outline for an Interrogative/Prompt Library to help improve output quality from Generative-AI Datasets. *Prompt Library to help improve output quality from Generative-AI Datasets (May 2023)*.

Turpin, M., et al. "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting." *Advances in Neural Information Processing Systems* 36 (2024).V1 submitted before July 2023 on open archive

Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022, April). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In CHI conference on human factors in computing systems extended abstracts (pp. 1-7).

Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022, April). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In Chi conference on human factors in computing systems extended abstracts (pp. 1-7). Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design Principles for Generative AI Applications. arXiv preprint arXiv:2401.14484.

Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, 614(7947), 224-226.

Van Maanen, J. E., & Schein, E. H. (1977). Toward a theory of organizational socialization.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023a). Explanations can reduce overreliance on ai systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-38.

Vasconcelos, H., Bansal, G., Fourney, A., Liao, Q. V., & Vaughan, J. W. (2023b). Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. arXiv preprint arXiv:2302.07248.

Waardenburg, L., Huysman, M., & Sergeeva, A. V. (2022). In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. Organization science, 33(1), 59-82.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526-1541.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214-229).

Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). Toward general design principles for generative AI applications. arXiv preprint arXiv:2301.05578.

Xu, F., Vasilescu, B. and Neubig, G. (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]

Yang, Q., Suh, J., Chen, N. C., & Ramos, G. (2018, June). Grounding interactive machine learning tool design in how non-experts actually build models. In Proceedings of the 2018 designing interactive systems conference (pp. 573-584).

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April). 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 (pp. 1-21).

Zetka, J. R. (2003). Surgeons and the Scope. Cornell University Press.

Table 1: Similar vs. New Sources of Risk Associated with Juniors Coaching Seniors in Emerging Technologies

	TECHNOLOGIES STUDIED IN EXISTING LITERATURE	EMERGING TECHNOLOGIES
SIMILAR SOURCES OF RISK		
Historical relationship between positional and practical expertise	● Position and title corresponded with knowledge and skill, which was cumulative across roles	Similar
Disruption of new technology to existing knowledge and skill by distancing senior members' from their work	● Adopting new technologies may mean replacing embodied methods with teleoperated methods ● Analytical technologies may bring a a shift from direct, intentional analysis to distanced, encoded analysis	Similar
Produces results that are partly beyond the control of the user organization	● Technology may gather data directly from external sources, automatically calculate key metrics, and feed these calculations directly into organizational tools	Similar
Output is often difficult to understand	● It is often difficult for users to comprehend how the technology works	Similar
Data may lead to biases in decisions of AI systems	● Technology may be trained on imbalanced data that lead models to exhibit, for example, a higher error rate for some demographic groups than others	Similar
NEW SOURCES OF RISK		
Simplicity and infrastructure required for access and usage by novice users	Interacting with AI requires infrastructure and raises barriers for ordinary users	The technology can be accessed and customized by novice users without coding and without owning infrastructure
Range of applications	A single system is often purpose built to execute a particular task	A single system can perform a broad range of applications
Rate of change	The technology is quickly evolving through product life cycle enhancements and user modifications	The technology is changing at an exponential rate
Has possibility of outperforming humans in a wide variety of skilled and cognitive tasks	Technology helps to inform and supplement human action	Technology has unprecedented, often superhuman performance to act more autonomously and conduct more tasks
Source of data for the technology	Data used in the system is often generated within a company	Data used in the system stems increasingly from a mix of various entities that dynamically interact

Table 2: Novice AI Risk Mitigation Tactics

	NOVICE AI RISK MITIGATION TACTICS	EXPERT AI RISK MITIGATION TACTICS
GENAI CAPABILITIES	**NOVICES MAY LACK DEEP UNDERSTANDING OF GENAI**	**EXPERTS ARE MORE LIKELY TO HAVE DEEP UNDERSTANDING OF GENAI**
Accuracy	● Use a standardized way of asking questions ● Do the work first without GenAI	● Decide on appropriate use cases where error risks are acceptable ● Test GenAI's reliability in executing each subtask
Explainability	● Explain model logic to managers ● Agree on practices for explainable output	● Avoid GenAI use where high degree of explainability is required ● GenAI provides illusion of transparency, but explanations may not match true actions
Contextualization	● Use for cases where contextualization is not necessary	● Provide contextual information, and specify the desired output ● Use RAG to add content
TARGET OF CHANGE	**HUMAN ROUTINES**	**AND DATA, MODEL, SYSTEM DESIGN**
Accuracy	● Train users to validate results ● Managers review user prompts/responses	● Set up second automatic monitoring system to check if in line with users' goals ● Use model that provides link to sources ● Use a more accurate model
Complacency	● Train users to take ownership of work when using GenAI	● Design a system that provides proactive self-reflective prompts ● Build an interface that visualizes uncertainty ● Build a default prompt that prompts the user to input their goals; Support pattern-matching between GenAI suggestions and users' task goals ● Apply hard-coded restrictions
Contextualization	● Train users in prompt engineering (also expert tactic) ● Gain agreement re: don't require contextualization	● Improve prompts centrally and build them into the system ● Design system to begin with a user prompt to communicate goals
INTERVENTION	**PROJECT LEVEL**	**AND SYSTEM-DEPLOYER AND ECOSYSTEM LEVEL**
Accuracy	● Managers and users on project agree on conditions under which GenAI can be used reliably ● Managers review users' work process	System-Deployer-Level ● Provide co-audit tools ● Communicate to users the intended conditions under which GenAI can be used reliably ● Create a prompt library of effective prompts for particular tasks ● Continually assess the alignment of LLMs vis-à-vis evaluation metrics ● Establish feedback/incident reporting mechanisms Ecosystem-Level ● Assess credibility of data sources; use trusted sources; Perform real-time data updates ● Flag/correct misleading responses ● Explain system's capabilities and limitations ● Avoid anthromorphizing the AI
Complacency	● Managers give users adequate time ● Managers don't shortsell cases	System Deployer-Level ● Effectively onboard users; Provide personalized adjustments for users ● Apply cognitive forcing functions ● Allow users to clearly specify human-GenAI task allocation Ecosystem-Level ● Attend to the impact on users of a model's style, tone, or perceived personality ● Highlight prompt changes and the resulting output changes ● Support debugging by generating test cases and test data to help identify corner cases

APPENDIX

Table A: Juniors Expected that they Would need to Educate Seniors, but did not expect Status Threat to be a Key Barrier

	EXAMPLES FROM THE INTERVIEWS WITH JUNIORS
Juniors expected to educate seniors about effective use of the new technology	**Juniors expected to educate seniors about effective use of the new technology** PS36: "We'll need to teach managers to accept our use of GenAI…[So], you need to try to educate them, and to alleviate some of their concerns." PS38: "You can start by picking the areas that [the managers] do feel comfortable in. For example, like fact aggregation, or whatever, and begin to implement [GenAI] into our ways of working [there], and couple that with learning labs to test it." PS21: "I would probably start using AI for the internal portion of our work, [like team communication], so that the managers can get a sense of how things improve with [the use of GenAI]." PS58: "Managers may differ in their acceptance of it… Potentially there are some trust issues that need to be addressed… But I would be OK altering my approach based on their preferences. That could also be mitigated by us just exposing them to it more." PS66: [Because my managers are supportive], when I discover something cool, I will share it with my managers. Like the new [GenAI] PPT tool. PS31: ""I think the biggest thing for managers is us just training them….We need to show managers that [GenAI] is not a magic wand waving solution. Maybe having them even just watch how an associate would do a single module, or have an understanding of what that looks like, would help them kind of right size in their head what [GenAI] can do….I think they have very little visibility is what I've seen from my side."
No evidence that juniors expected to need to mitigate status threat to seniors that this upward teaching could raise	**Juniors did not expect to need to teach seniors without appearing to do so, in order to protect seniors' status** PS36: "Put the reasoning in front of their faces." PS38: "It's just about creating a dialogue with them about what you are doing. And it's about making them feel more comfortable with it, through trial and error and trainings." PS39: "I think making the case to the hesitant manager will go a long way, like, 'You know, it's going to save me time this in this way. And that gives me opportunity to work on these other aspects of the project that are super important."

Table B: Juniors Expected the Key Barrier to Upward Coaching Would Be the Technology's Risks to Seniors' Valued Outcomes

RISK TO VALUED OUTCOMES	EXAMPLES FROM THE INTERVIEWS WITH JUNIORS
Accuracy of Outputs	PS24: "Obviously the first [thing managers will worry about is that] leveraging GenAI can produce output that is actually incorrect. So that would be a big concern." PS3: "Managers will be scared that [new consultants] would bring something made up…If a [consultant] is doing desk research on new technology around vitamins and minerals, he could ask GPT, and ask for the links. Since he doesn't know how consulting works, he might use the links directly from GPT, and give it to the manager." PS20: "Managers will be afraid of AI coming up with answers, and junior people will think know whole world, and will talk to clients which will discount our credibility with them." PS32: "I imagine most managers would probably steer us away from using it, because they don't want to put their reputation at risk if things turn out to be wrong." PS13: "[Managers will be concerned that the client] might take some of our results and do stuff with it by hand, and realize that we used Generative AI in a way we weren't supposed to. Cause an Associate was tired at night." PS59: "Managers may worry that more junior people may over-trust GenAI, and provide wrong data or answers to manager. Then, if the manager doesn't check, they may advise client incorrectly." PS60: "I can see my [manager] having an issue with [consultants using GenAI], because they always want the end result to be correct and AI can be wrong."
Explainability of Outputs	PS6: "Using GenAI will [negatively] affect a lot of the communication between consultants and managers, because consultants used to be able to back up everything, and now they won't be able to." PS50: "When [managers] don't understand the tech, they find it hard to relate to it and trust it…We need to decide how consultants should explain the results from AI [to managers]." PS7: "Managers will be worried about GPT [because] it's a black box, and there's no way to check if it is accurate."
Contextualization of Outputs	PS24: "[Managers will] be concerned about the tool being able to take into account all the variables or context that are actually necessary to put together a valid solution…It's all about context. If you are working in a very nuanced industry with a very specific set of guidelines for the client, having that knowledge is one thing, and being able to feed that into a tool to actually spit out valid or actionable recommendations is another." PS63: "Managers may be concerned about, 'Is GPT output specific enough [given the context of the case and the client's problem]?'…Like, if it's for a client that has a lot of their financial information that isn't available online, [then managers would be concerned about] is [GPT output] being tailored and specific enough to the client's context as opposed to the generic publicly available information." PS12: "Managers will worry about consultants' lack of context because GPT is very focused on what you just fed it."

Table B: Juniors Expected the Key Barrier to Upward Coaching Would Be the Technology's Risks to Seniors' Valued Outcomes (continued)

RISK TO VALUED OUTCOMES	EXAMPLES FROM THE INTERVIEWS WITH JUNIORS
User Engagement	PS12: "Managers will worry about complacency. They will want to make sure that consultants aren't just blindly accepting GPT output." PS13: "What will be difficult for managers…is not micromanaging, and trusting the results provided by especially new team members…[managers will] have a little bit of worry that some people are just ripping into generative AI and passing it along." PS5:" I think managers wouldn't want to get into a situation where anyone is overly reliant on AI." PS20: "Managers' main concern would be that consultants are not blindly believing like everything the tool spits out…there will be people who put zero to 5% of human touch and rely super heavily on AI." PS73: "If I were a [manager], and I learned that the analysis was done using GenAI, I would have a couple of concerns. One is, was there any vetting of the quality of the of the output?" PS63: "I think there will be a trust issue, [with managers wondering], 'Did you do this yourself, or did you use a tool to do it? Did you really think through that it was the right approach? Have you double checked it? Did you reread it through, and edit it to better fit our context? Did you look at any other sources to check if this is accurate information?' All those questions of, did you do your due diligence on this work that someone else effectively did for you." PS64: "If there was a lazy AC on the team, and they used AI for everything and didn't bother quality checking, refining it. [Using GenAI] could let them get away with poor quality work. Before AI, you would still have had to come up with something yourself." PS70:" I think there's going to be a trust disconnect to a degree, where managers don't believe that the consultants are fully involved in building this thing…consultants may [start] shutting off earlier, and getting to a 90% answer. I think there's going to be a trust issue that I'd very much envision especially if it's [consultants using it for critical] points in the analysis, like if someone's using it [for certain things] like building model outputs."

Table C1: Examples of Novice vs. Expert AI Risk Mitigation Tactics

	NOVICE AI RISK MITIGATION TACTICS	EXPERT AI RISK MITIGATION TACTICS
GENAI CAPABILITIES	1. **NOVICES LACKED A DEEP UNDERSTANDING GENAI CAPABILITIES**	**EXPERTS HAD A DEEP UNDERSTANDING**
Accuracy	**Use a standardized way of asking questions** PS19: "We could have a standardized way for asking questions in GenAI, and a standardized way of summarizing the AI steps, and make that available to everyone." PS63: "Managers could give consultants] a framework of how they should approach [using GenAI], how they should coach the AI tool through that. Then [managers would] be more confident in it, than if consultants just give [GenAI] the question and let it answer it, however, it likes." **Do the work yourself first, before using GenAI** PS24: "GenAI tools should be used later on in the process. I think that when you're working on that initial problem-solving component, you should be doing it yourself, manually. And then once you have more of a finalized output, you can cross check, or use generative AI to be able to add on any kind of additional thoughts that you might have. So, we could leverage that approach. Do the work yourself first. Only use Gen AI once you have any kind of preliminary output, and augment only versus create."	● Decide on appropriate use cases where error risks are acceptable (OpenAi, 2023a) ● Test GenAI's reliability in executing each subtask (OpenAI, 2023b)
Interpretability	**Explain GenAI logic to seniors** PS25: "We should…explain its rationale… We have to be able to do a deep dive and explain everything." PS52: "The manager may question what GenAI did, so you need to be able to explain it." PS41: "It's about understanding the source of the recommendation or the result. Being able to explain it." **Agree on practices for explainable output** PS19: "You could be careful in asking the AI how it got to the answer, not sure how well it works. We could have a standardized way of summarizing the AI steps." PS47: "We need to align on what the process is for using [GenAI], and how the output should look. There are some tools that we use a lot…where we are quite aligned on how the output should look. So maybe there needs to be a similar kind of thing with [Gen]AI where there becomes [our organization's] way of using it, and showing it."	● Avoid GenAI use where explainability is required (Bender et al., 2021; Liu et al., 2023) ● Provide user with global explanations about model logic and how to improve the input (Liao and Vaughn, 2023

Table C1: Examples of Novice vs. Expert AI Risk Mitigation Tactics (continued)

	NOVICE AI RISK MITIGATION TACTICS	EXPERT AI RISK MITIGATION TACTICS
GENAI CAPABILITIES	1. **NOVICES LACK A DEEP UNDERSTANDING OF GENAI CAPABILITIES (continued)**	**EXPERTS HAVE A DEEP UNDERSTANDING**
Contextualization	**Use for cases where contextualization is not necessary** PS20: "We could implement AI without context in limited areas, where it's fine to have generic answers. So, places like knowledge research, industry trends, slides creating transactional or tactical words that require minimal human touch. Wherever a generic answer is applicable, it's fine for consultants to say, 'I searched using AI.'" PS63: "If you're using it for [a case] that's based off stuff you can search on the Internet, then maybe [using GenAI] makes sense. Like for a project that is sort of generic and in the public domain. I did a project on wildfire policy that probably would have been great to use [GenAI] for, because we spent a lot of time digging up generic information that's available online."	● Provide contextual information, and specify the desired output (Zhou et al., 2023) ● Use RAG to add content (Lewis et al., 2020; Gu et al., 2020)

Table C2: Examples of Novice vs. Expert AI Risk Mitigation Tactics (continued)

	NOVICE AI RISK MITIGATION TACTICS	EXPERT AI RISK MITIGATION TACTICS
TARGET OF CHANGE	2. **CHANGE HUMAN ROUTINES**	**And CHANGE DATA, MODEL, SYSTEM DESIGN**
Accuracy	**Train users to validate results** PS15: "I think managers would struggle if consultants began to rely on GPT too much….And, I can see some consultants copy pasting, too, and getting inaccurate results. To mitigate this, we should train people on how to coach GPT to do certain things." PS76: "We should have content sessions where we really drill consultants on questions to get second order insights, and check that they really understand, and are not just relying on the technology." PS63: "[It will be important to teach consultant to] do spot checking of the numbers so that they have confidence in how [GenAI] did the approach, and [teach them to do] the critical thinking of, does the approach and does the answer it gave me makes sense. And to think about sources and actually going and digging up each of those sources to make sure they are real, and that is what that source is saying." **Have managers review prompts/responses** PS7: "Consultants can handle manager concerns by showing managers the prompts they provided to GPT." PS12: "It would be good to show the manager what you put into GenAI to get that result, because then managers know for sure what you fed to the AI tool before you generated the output. That would help make sure that you didn't get any wrong information, and would give the manager a record of how you got that, to make sure it's not crazy stuff." PS26: "Maybe there could be a way of including the prompts you sent AI or some way for managers to check that part…Consultants can say, these are instructions I gave, and managers can tell us if there are any issues with those instructions." PS56: "If you did use it, it could be difficult to cross-check what you did. [Providing our managers with] chat transcripts and retracing our steps of analysis could help. As long as you have ways to collaboratively validate [the output], it would be fine."	● Set up second automatic monitoring system (Saunders, 2022) ● Use model that provides link to sources (e.g., Simkute et al., 2024) ● Use a more accurate model (Win et al., 2023)

Table C2: Examples of Novice vs. Expert AI Risk Mitigation Tactics (continued)

	NOVICE AI RISK MITIGATION TACTICS	EXPERT TACTICS
TARGET OF CHANGE	2. **CHANGE HUMAN ROUTINES (continued)**	**And CHANGE DATA, MODEL, SYSTEM DESIGN**
Complacency	**Train users to take ownership** PS27: "Consultants need to know that the ownership is theirs, at the end of the day. If they use AI or not, they are presenting this answer; it is on them. The excuse can never be, 'I put this in AI, it's a black box.' They have to understand it. Consultants can't siphon the responsibility off. They own that answer." PS45: "I think there's a feeling sometimes among managers that, oh, if my associate consultant is using generative AI, they're just literally putting everything in the chat GPT and copying and pasting what it says. So managers need to know how to make sure that consultants have an idea of what it takes to get the best output from generative AI. So that if [managers] do see there is a consultant who is just copying and pasting everything, [the manager] can push them to be like, "No. You need to get more out of this." PS26: "One big issue is consultants need to learn how to QA the output to take ownership of their work [with GenAI]." PS23: "AI is not this invaluable tool; it is still your decision. [Consultants] are not in a position to say, "AI did this. They need to learn that it is still their decision, their data they're presenting."	● Build an interface that visualizes uncertainty (Vasconcelos et al. 2023b) ● Support pattern-matching between GenAI suggestions and users' task goals (Barke et al., 2023) ● Apply hard-coded restrictions to improve an LLM's alignment to a custom objective (Lu et al., 2022)
Contextualization	**Train users in prompt engineering** PS4: "[We need] training on how to better structure prompts. How to make sure that you are giving cues in terms of tone, audience, things like that. As well as if you are looking for a specific level of detail in certain areas and how to structure the output that you want." PS15: "We should train people on how to coach GPT to do certain things by teaching them techniques like prompt engineering." PS43: "Give people examples of prompts they can use…what the related outcomes are." **Gain agreement within the team about which uses don't require much contextualization** PS31: "We just need to get agreements within the team…we can all agree that a 90% version of a document shouldn't be GPT, because in consulting, we come up with creative solutions. Every client needs a unique solution, and that's hard to do with GPT template."	● Align the model's generations to particular objectives specified by users by using human-labeled preference data at the fine-tuning stage (e.g., Song et al., 2023) ● Design system to begin with a user prompt to communicate goals (Zamfirescu-Pereira et al., 2023) ● Improve prompts centrally and build them into the system (Achiam et al., 2023)

Table C3: Examples of Novice vs. Expert AI Risk Mitigation Tactics (continued)

	NOVICE AI RISK MITIGATION TACTICS	EXPERT TACTICS
INTERVENTION LEVEL	3. **INTERVENE AT PROJECT LEVEL**	**And INTERVENE AT FIRM, ECOSYSTEM LEVEL**
Accuracy	**Seniors and juniors on a project agree on conditions under which GenAI can be used reliably** PS70: "I imagine a world where a manager would have to set very clear boundaries up front…I think it'll have to be a very clear discussion of, here's my expectation as a manager of when you use it and when you don't. I think it'll be a burden on the manager to set that versus the consultant having to be unsure of when it's okay when it's not okay." PS65: "At the beginning of the project, we just need to set the guidelines around what to use it for and not and why." PS44: "Managers need to set that expectation explicitly at the start of the case about how GenAI can be used. You're going to run into issues when you don't define those guardrails clearly, and then you'd potentially run into conflicting expectations in the middle of the case that managers probably want to avoid." PS10: "We'll need to come to agreement within the team around to what degree do we want answers from AI. Should we use it only for first ideas research, or for full analysis." PS13: "What managers could do is specify what they allow [GenAI] to be used for and how to check its use. Managers need to put safeguards around how they want it used." **Managers review consultants' work process** PS5: "Managers will need to ask consultants, 'What did you put in [to GPT]? How did you structure your analysis?" PS11: "If I were a manager now, I would set 30 min for each analysis, and tell the consultant to walk me through how they did it with GenAI." PS13: "I think the biggest thing [with GenAI is]…having the manager do what managers currently do, which is sporadically ask questions about how analyses were done. So managers…make sure that the consultant actually did it." PS70: "I think it puts an extra level of burden on managers to validate the results. Right now, it's one level [of validation]. A consultant is doing all this work to then validate it…[With GenAI, managers could be concerned] that consultants will start to cut down time on those tasks, because they're using generative AI, and they're doing [only] some level of validation. [Managers may] think that [consultants] are not going to as deeply understand the nuances of the problem, bugs, things that. So, a manager is going to ask questions or try to validate it themselves." PS72: "Managers might think that it's not fully our ideas with ChatGPT…[To mitigate this] you could try using it together. Sit down and assess answers together and iterate with it." PS77: "Managers will need to be even more skeptical of the results and drilling a bit harder. Because they will wonder about whether or not I checked my work."	**System Deployer-level** ● Provide co-audit tools (Mündler et al. 2023) ● Communicate to users the intended conditions under which GenAI can be used reliably (Liao and Vaughn, 2023) ● Create a prompt library of effective prompts for particular tasks (e.g., Svendsen et al., 2023) ● Develop evaluation methods that take into account emergent capabilities as models get more capable, and are open ended enough to detect unforeseen risks (OpenAI, 2023a) **Ecosystem-level** ● Explain system's capabilities and limitations (Solaiman et al, 2023) ● Build in effective explanations (Vasconcelos et al., 2023a) ● Assess credibility of data sources; use trusted sources (Li et al., 2023) ● Adjust LLM behaviors by implementing algorithms to assess the recency and applicability of data points (Meng et al., 2021)

Table C3: Examples of Novice vs. Expert AI Risk Mitigation Tactics (continued)

	NOVICE AI RISK MITIGATION TACTICS	EXPERT TACTICS
INTERVENTION LEVEL	3. **INTERVENE AT PROJECT LEVEL (continued)**	**And INTERVENE AT FIRM, ECOSYSTEM LEVEL**
Accuracy (cont.)	**Tailor use according to degree of manager acceptance** PS38: "I think different managers are going to have different perspectives on how much GPT should be used, and what it should be used for and not. Some will be very supportive of it, others will not…You need to follow their lead…and use it to the degree they feel comfortable. You can begin to leverage it if they do, but if they don't, you shouldn't." PS46: "Of course, managers will have their own perception. So, consultants should use it to the degree the manager is comfortable with. And [consultants] should overcommunicate every step." PS48: "I would adapt based on my manager's perception of it." PS21: "If there is a skeptical [manager], I would probably start using AI for the internal portion of our work, [like team communication], so that the manager can get a sense of how things improve with [the use of GenAI]."	
Complacency	**Managers give end users adequate time** PS42: "Frequently we are under time pressure and that leads to not being able to quality check as much as you'd like. So, ideally, managers would need to give more time for quality checking. I think that would be important." PS76: "We need to train managers to give consultants enough time to do the work. If my manager does not give me enough time, if I have to choose between doing my work and doing normal human things, I will choose to do human things." PS57: "One challenge is that using GenAI would increase speed of operations leading to increased expectations from managers, but what if [the consultant is] not confident in the output? To mitigate that, we'd need to train managers to give extra time for quality checking." **Managers don't shortsell cases** PS76: "[Once managers] start thinking that because consultants have [GenAI] they can do things faster, managers will have different expectations…Managers will start to expect this time savings and sell projects based on it. Then, consultants will stop getting details, and will need to compromise on their depth of understanding… Consultants are already working at a high level. We will begin to learn things in one week and think we are experts. If managers don't give consultants the time, I fear that managers will short sell, and consultants will compromise…Tell manager not to overscope and shortsell a project." PS49: "It could mean that managing [managers'] expectations in terms of output from the team and the number of the people on the team becomes a lot more important." PS18: "It will be really necessary to manage [managers'] expectations of what they can receive, if we use Gen AI versus doing it as we've done traditional consultant work."	System Deployer-level ● Effectively onboard users ; Provide personalized adjustments for users (e.g. Lu & Yin 2021) ● Apply cognitive forcing functions (Buçinca et al. 2021) ● Allow users to clearly specify human-GenAI task allocation (Simkute et al., 2023) Ecosystem-level ● Highlight prompt changes and the resulting output changes (Zamfirescu-Pereira et al., 2023) ● Support debugging by generating test cases and test data to help identify corner cases (Vaithilingam et al., 2022)