Beneficial Mistrust in Generative AI: The Role of AI Literacy in Handling Bad Advice

Authors blinded for review[1*]

[1]Blinded for review

*Corresponding author: Blinded for review

Despite the increasing proliferation of Generative Artificial Intelligence (GenAI), systems like large language models (LLMs) can sometimes present misleading or false information as true – a problem known as "hallucinations." As GenAI systems become more widespread and accessible to the general public, understanding how AI literacy influences advice-taking from imperfect GenAI advice is crucial. Drawing on the correspondence bias, we study how individuals with varying AI literacy levels react to GenAI providing bad advice. Gathering empirical evidence through an online programming experiment, we expect AI-literate individuals to take less advice *while* receiving an error but relatively more advice in *future* tasks. We outline how correspondence bias can explain these variations, reconciling mixed findings of prior studies on AI literacy. Our research thus contributes a holistic perspective on the beneficial mistrust through AI literacy to education, integration, and evaluation programs of AI, highlighting the dangers of naive evaluation strategies.

Keywords: *AI literacy, Artificial intelligence, AI education, Algorithm aversion*

1 Motivation

Advances in Artificial Intelligence (AI), particularly Generative AI (GenAI), are fueling its increasing use among programmers and knowledge workers (e.g., Surameery and Shakor 2023; Dell'Acqua et al. 2023). GenAI refers to AI systems that create new content, such as text, images, or code, by learning from existing data (OpenAI 2024). However, these systems are not flawless and can produce misleading or false information, known as "hallucinations" (Maynez et al. 2020). When encountering such "bad advice," users may attribute it to situational factors (e.g., insufficient context in the prompt) or dispositional factors of the GenAI (e.g., model limitations). Such attributions are crucial as they influence advice-taking both while receiving bad advice (task t) and in future tasks (task t+1).

AI-based systems have become more user-friendly, making them accessible not only to AI experts but also to individuals from different backgrounds (Maedche et al. 2019). As a result, the literature suggests that individuals require new competencies to effectively engage with AI, commonly referred to as AI literacy (Long and Magerko 2020). AI literacy refers to a "set of competencies that enables

individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (Long and Magerko 2020, p.2). Enhancing individuals' AI literacy enables them to align their evaluation of AI's appropriateness with their delegation of decisions to AI (Pinski et al. 2023). For instance, understanding the rationales behind AI output as one competency linked to AI literacy is crucial for appropriate interaction and dealing with the inscrutability of AI (Asatiani et al. 2021; Lebovitz et al. 2021). Therefore, gaining a deeper understanding of individuals' interactions with AI-based tools is imperative to mitigate potential negative outcomes in case of bad advice from an AI (Benbya et al. 2021).

However, there are mixed results regarding the direct effects of AI literacy on advice-taking. Developing the ability to question, explain, and understand an AI's activities is crucial for building trust in AI (Martin 2019), which is positively related to advice-taking. Yet, AI literacy can also reduce individuals' willingness to use AI in the future (Pinski et al. 2023). Individuals with a limited understanding of AI are more likely to overestimate the intelligence of AI (Ehsan et al. 2021) due to unrealistic expectations (Dietvorst et al. 2015), which can also increase advice-taking. Thus, we aim to investigate how AI literacy influences advice-taking.

While expectations in GenAI grow, particularly for enhancing productivity in domains such as programming (Peng et al. 2023), AI is not free from errors and can create bad advice (Lebovitz et al. 2021). In contrast to prior technologies, detecting and fixing bad advice from GenAI, such as hallucinations, remains an ongoing research frontier (Berente et al. 2021; Ji et al. 2023). Thus, we are interested in advice-taking while receiving bad advice (task t). Blindly following advice can be problematic if an AI's advice fails to meet the users' expectations (Stieglitz et al. 2023). As with any setback, individuals naturally seek to understand the reasons behind the event (Heider 1958). Bad advice attributed to the competencies of the advisor can lead to a quick reputation loss of the advisor and thereby decrease future advice-taking (Yaniv and Kleinberger 2000). Therefore, we are also interested in how individuals' AI literacy influences future advice-taking (task t+1) due to different causal attributions of a GenAI's hallucination. Thus, we aim to answer the following research question:

What is the influence of AI literacy on individuals' advice-taking behavior while receiving bad GenAI advice and their advice-taking behavior in future tasks?

To address our research question, we build on the correspondence bias, suggesting that individuals attribute bad advice to dispositional factors (e.g., GenAI's model) rather than situational factors (e.g., forces external to the GenAI such as a prompt missing context details) (Gilbert and Malone 1995). We will conduct an online programming experiment, extending the quantitative results with qualitative insights on the correspondence bias (Reis et al. 2022). First, we will investigate individuals' advice-taking while receiving bad advice from GenAI (task t) and how they take advice in the next task when receiving new advice (task t+1) using the Judge Advisor System (JAS). Thereby, we investigate two factors: the advice quality (bad vs. good advice) and individuals' level of AI literacy. Second, we will qualitatively analyze open-ended questions to explore advice-taking differences across AI literacy levels. We expect bad advice to decrease advice-taking. In addition, we anticipate AI literacy to mediate this decline, with AI-literate individuals taking less advice while receiving bad advice and relatively more advice in future advice-taking when receiving new good advice compared to less literate individuals.

We anticipate to contribute to three key areas of research. First, we provide empirical evidence that AI-literate individuals take less advice *while* receiving bad advice but relatively more advice in *future* tasks when advice quality is good. Second, we propose correspondence bias as an explanation for differing advice-taking behaviors, suggesting that AI literacy moderates the impact of bad advice on advice-taking by mitigating correspondence bias. Third, we adopt the Judge-Advisor System (JAS) framework from psychology and apply it to a programming task supported by GenAI — a context of growing practical relevance. Complementing our quantitative findings, qualitative insights reveal why individuals with varying AI literacy levels follow or reject GenAI advice. This holistic approach informs educational programs, integration strategies, and AI evaluations, emphasizing the critical role of AI literacy in fostering beneficial mistrust in GenAI.

2 Theoretical Background: The JAS & The Correspondence Bias

To evaluate the extent to which individuals incorporate guidance from GenAI into their own decision-making process and their rationale behind it, we utilize the Judge-Advisor System (JAS) as a framework. In the JAS, a person referred to as a "judge" faces a decision scenario and forms an initial judgment. Subsequently, the judge receives advice from an advisor, such as a GenAI. The judge then combines his or her initial judgment with the advice received from the GenAI, making sure to appropriately adjust the initial judgment and weigh the advice to arrive at a final judgment (Bonaccio and Dalal 2006).

Judging the advice to arrive at a final decision can be subject to a certain bias. The correspondence bias describes the tendency of humans to attribute observed behavior to dispositional factors (i.e., reliability, ability, or skills) while underestimating the role of situational influences such as the environment or situation (Gilbert and Malone 1995; Edwards and Edwards 2022). It is part of attribution theory, which seeks to explain how individuals interpret the reasons behind certain observations categorizing them as either situational factors or dispositional factors (Kelley and Michela 1980).

Individuals' attribution process is influenced by the perception of the situation, the behavioral expectations, and the perception of the observed behavior (Gilbert and Malone 1995; Edwards and Edwards 2022). Correspondence bias occurs because individuals rely on their causal theories about how situations influence behavior when making judgments about others. In general, individuals tend to attribute behavior to characteristics or dispositional factors of a person, as it is easier and faster for them to access than evaluating situational dimensions (Gilbert and Malone 1995).

Although correspondence bias was first observed in human-to-human interactions, there is a growing tendency for individuals to perceive AI not merely as a tool but as a collaborative teammate (O'Neill et al. 2022), often forming emotional bonds in the process (Müller 2021). Hence, preliminary research has begun to extend the application of correspondence bias to the context of human-AI interactions (Edwards and Edwards 2022), as individuals increasingly respond to AI and humans in similar ways by attributing responsibilities to AI (Hong et al. 2020).

Figure 1 illustrates how insufficient context can create hallucinations, leading to bad advice. It also shows different ways correspondence bias can influence individuals' attribution of responsibility throughout this process. On a technical level, GenAI tools supporting coding processes, such as GitHub

Copilot, access the context of a programming task through code and data shared with the GenAI (e.g., Salva 2024). However, the prompt given to the GenAI (i.e., code written by the user) may lack contextual information (i.e., data columns of a file not shared with the GenAI). Consequently, GenAI has to rely on its inherent model to provide coding advice and suggest the most likely column names based on the available code. These predicted column names may not exist in the actual data, leading to hallucinations causing bad advice. Therefore - in this study's context - situational factors (such as the prompt lacking contextual details) can entirely explain the bad advice, as the inherent model of the GenAI correctly provided the most likely suggestion, even though it was bad advice in this situation.

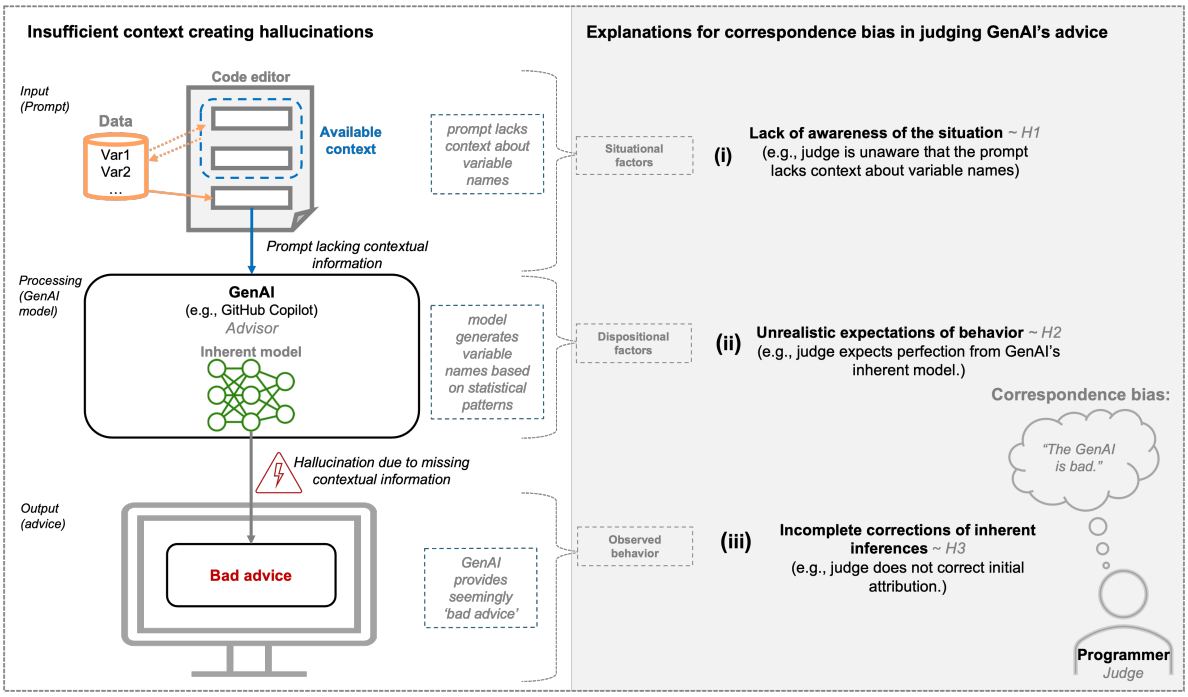

Figure 1: Insufficient context creates hallucinations and may cause correspondence bias.

The programmer (judge), on the receiving end of GenAI's advice, can only observe its performance and seek to identify potential causes and explanations for instances of bad advice. Those explanations are likely to shape the programmer's behavior while receiving bad advice (task t) and in future advice-taking scenarios (task t+1). Prior research suggests that the correspondence bias can shape such attribution processes, leading individuals to overemphasize dispositional factors (e.g., the quality of the GenAI model) as the primary cause of the observed behavior (e.g., bad advice), while underestimating situational factors (e.g., contextual limitations in the prompt) that may have also influenced the outcome.

Following Gilbert and Malone (1995), three established explanations contribute to the phenomenon of correspondence bias, as illustrated on the right side of Figure 1: lack of awareness of the situation (i); unrealistic expectations of behavior (ii); and incomplete corrections of inherent inferences (iii). Those three explanations will guide the hypothesis development (see Section 3) and the derivation of the qualitative questions (see Section 4.4). These explanations provide insights into and explanations for differences in advice-taking behavior among individuals possessing varying levels of AI literacy, as we expect higher levels of AI literacy to mitigate the correspondence bias. We outline the psychological foundation of the correspondence bias in the following:

(i) **Lack of awareness of the situation** refers to individuals overlooking contextual circumstances influencing a GenAI's advice due to an incomplete understanding of how such circumstances affect the GenAI's functioning. While situational circumstances - such as data columns of the file being inaccessible to GenAI - significantly influence GenAI's advice, attributing bad advice to situational factors requires an awareness of their existence. Previous studies have noted that when judges are uncertain about the reasons behind an advisor's behavior, they often attribute it to dispositional factors of the advisor (Gilbert and Malone 1995). For instance, organizational employees attributed the cause of critical incidents in human-computer interactions to failures of advisors, such as software, in 74% of cases (Ortiz de Guinea 2016). In the context of GenAI, causes for critical incidents could be situational factors such as missing context details in the prompt that lead to hallucinations.

(ii) **Unrealistic expectations of behavior** implies that judges can experience correspondence bias if a GenAI's behavior deviates from their expectations. A study by Ross et al. (1977) illustrates this phenomenon in human-human interaction. They asked participants whether they would wear a sign with a certain inscription. Those willing to wear a signboard expected others would also be willing. Thus, they attributed the declining behavior of others to their dispositional factors, such as their political attitude. Similarly, judges often expect perfection from algorithms such as GenAI (Dietvorst et al. 2015). Specifically, if they overestimate the GenAI model, bad advice can lead to correspondence bias. More precisely, not knowing the GenAI model's functioning and its related limitations may increase the tendency to attribute bad advice to GenAI's dispositional factors.

(iii) **Incomplete corrections of inherent inferences** implicates that judges may lack either the motivation or the ability to adjust dispositional inferences they spontaneously and effortlessly form (Gilbert and Malone 1995). For example, users of conversational agents attributed agency and intention to humanlike GenAI (Brendel et al. 2023). Then, they may lack the motivation to correct the initial attribution of dispositional factors to the GenAI model. As illustrated in Figure 1, individuals who face bad GenAI advice in one task may lower their expectations of GenAI and may continue to hold these expectations even when GenAI provides good advice in the next task.

There are first investigations into correspondence bias in socio-technical settings. These studies indicate a shifting perception of AI, with it being increasingly viewed as a "teammate" rather than just a tool (Zhang et al. 2021), and attributed with certain human-like qualities (Brendel et al. 2023). This is evident by judges showing attribution bias in interactions with both humans and AI-based technology like social robots (Edwards and Edwards 2022). Overall, individuals seem to use "folk-psychological interpretations" when judging AI-based technology like robots (Thellman et al. 2017). We aim to contribute to current research by using the correspondence bias to explain individuals' behavior in JAS settings where humans have to solve programming tasks and receive GenAI-based advice. When GenAI produces hallucinations and individuals show one of the above-stated explanations, we posit that a correspondence bias may explain individuals' GenAI utilization.

3 Hypotheses Development

Figure 2 provides a conceptual overview of our research model, which examines the relations between bad advice, AI literacy, and advice-taking behavior. We will elaborate upon these hypothesized relations in the subsequent section by relying on prior empirical findings and the theory behind correspondence bias. Hypothesis 1a will examine the direct effect of bad advice (vs. good advice) on advice-taking for the assignment at task t. Hypothesis 1b will examine the effect of bad advice (vs. good advice) at task t for the next assignment at task t+1. Hypothesis 2 will examine the direct effect of individuals with higher AI literacy (vs. lower AI literacy) on advice-taking in task t. Hypothesis 3 will propose AI literacy moderating the effect of bad advice on advice-taking in both cases, task t and task t+1.

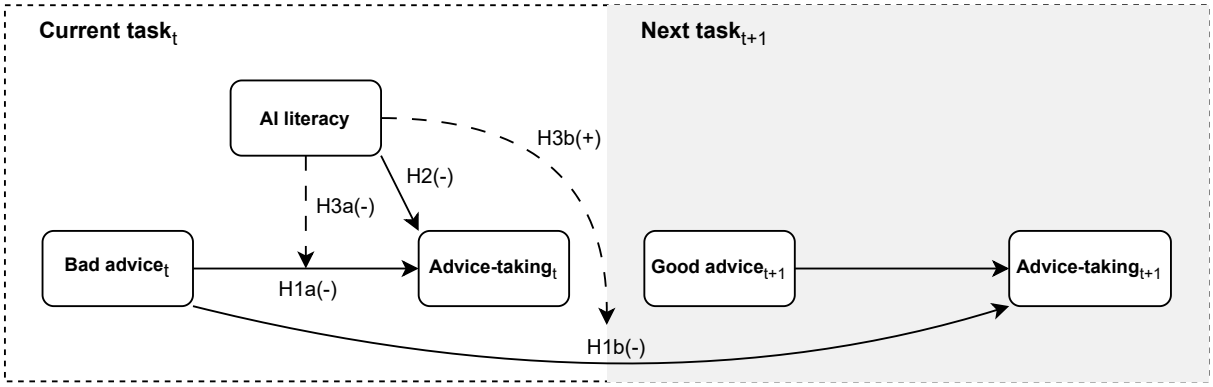

Figure 2: Research model including the three hypotheses for tasks t and t+1.

3.1 Bad Advice and Its Effects on Advice-Taking

Although research into GenAI has made significant progress in recent years, its probabilistic nature currently limits its ability to solely provide good advice. For instance, LLMs sometimes generate hallucinations, which remain an open field of research due to various causes for hallucinations (Ji et al. 2023). For instance, GenAI lacking file information can lead to hallucinations (e.g., Salva 2024). We ground our first hypotheses (H1a, H1b) in the expectation that some individuals have a *lack of awareness of such situational factors* (i).

Referring to H1a, we propose that individuals who receive bad advice are less likely to follow it than those who receive good-quality advice. There is both theoretical and empirical evidence that judges generally use good advice more than bad advice (see Bonaccio and Dalal 2006, for a review). For instance, when asked about dates of historical events, individuals use bad advice less, even in the absence of receiving feedback (Yaniv and Kleinberger 2000). Similarly, participants in a time series task on sales were able to distinguish between the reliability of statistical forecasts both with and without feedback (Lim and O'Connor 1995). When estimating the length of rivers, advice quality positively influenced advice-taking (Schultze and Loschelder 2021). Although a meta-study did not find a significant effect of high-quality advice on advice-taking compared to low-quality advice ($p = 0.122$), it indicated a positive correlation of 41%-points (Bailey et al. 2023). These results suggest that bad advice can affect an individual's willingness to take advice, leading us to hypothesize for task t:

Hypothesis 1a: Individuals receiving bad advice (vs. good advice) are less likely to take that advice.

A lack of awareness of situational factors can lead to correspondence bias not only *while* receiving bad advice (task t, H1a) but also in *future* tasks (task t+1, H1b). Individuals can attribute bad advice to dispositional factors, such as the advisor's perceived quality (Gilbert and Malone 1995). This attribution can damage the advisor's reputation, reducing future advice-taking (Yaniv and Kleinberger 2000). However, dispositional factors do not account for all instances of bad advice.

When individuals are unaware of situational factors, they exhibit correspondence bias in their attribution process (Gilbert and Malone 1995). Such a deterministic worldview opposes a probabilistic worldview, which considers incomplete information and accepts uncertainty (Einhorn 1986). In favoring deterministic explanations (Highhouse 2008), individuals, thereby, ignore the probabilistic nature of analytical approaches like GenAI. Instead of recognizing GenAI's probabilistic nature, they stigmatize its failures more than those caused by humans (Longoni et al. 2023). Additionally, they tend to attribute cognitive inflexibility to AI, assuming it cannot adapt to unexpected situations (Longoni et al. 2019). As a result, people exhibit stronger correspondence bias toward AI, such as robots, than toward humans (Edwards and Edwards 2022). Attributing bad advice to dispositional factors amplifies this bias and can reduce advice-taking.

There is evidence that advice-taking decreases future advice-taking (task t+1). For instance, individuals quickly lose trust when other humans provide bad advice (Yaniv and Kleinberger 2000). Instead of incorporating all information, individuals tend to base their attribution on recent events (Kahneman 2011). When faced with imperfect advice, individuals exhibited a stronger decline in advice-taking for algorithms compared to humans, a phenomenon known as algorithm aversion (Dietvorst et al. 2015). Studies on algorithm aversion observed a similar or stronger negative impact on advice-taking for algorithms, such as GenAI, compared to humans (Prahl and Van Swol 2017; Leffrang et al. 2023). Further studies observed that algorithm aversion only emerged after receiving feedback about bad advice (Berger et al. 2021). Focusing on the effect of observing bad advice on future advice-taking (task t+1), we hypothesize:

Hypothesis 1b: Individuals who previously received bad advice (vs. good advice) are less likely to take new good advice.

3.2 AI Literacy and Its Effects on Advice-Taking

Referring to our second Hypothesis (H2), we suggest that individuals with low AI literacy are more likely to take advice from GenAI due to *unrealistic expectations of GenAI's behavior* (ii) when judging GenAI's advice in JAS. AI literacy refers to the ability to reflectively interact with AI, including understanding its strengths and weaknesses (Long and Magerko 2020). This indicates that individuals with high AI literacy are generally more aware of how GenAI functions and the situational factors influencing its advice. Such awareness implies reduced correspondence bias. We expect an AI-literate person to know that hallucinations can occur if the AI lacks sufficient contextual information.

At first glance, AI literacy seems to increase advice-taking. In an experiment on AI knowledge as one aspect of AI literacy, individuals reading instructions about AI knowledge had a 14.1%-point higher

delegation rate to AI (Pinski et al. 2023). Similarly, explanations of AI advice led to an approximately 31%-point higher median Weight of Advice (Panigutti et al. 2022). However, in both studies, there is little evidence that the AI provided bad advice. On average, the AI improved the performance of the participants (Pinski et al. 2023) and the median perceived output quality of 3.2 on a 5-point Likert scale indicated no bad advice either (Panigutti et al. 2022). Evaluating advisors on unbalanced datasets, with mainly good advice, can bias the evaluation (see Jurafsky and Martin 2023, pp. 68-70).

In contrast, individuals with low AI literacy are more prone to overestimating the intelligence of AI (Dietvorst et al. 2015; Ehsan et al. 2021) and they tend to lack the expertise to fully comprehend the functioning of AI systems (Pinski and Benlian 2024). Overestimating an AI's ability may stem from a lack of experience in evaluating the shortcomings of AI advisors. Subsequently, unrealistic expectations of GenAI's behavior may trigger correspondence bias (Gilbert and Malone 1995).

Consequently, this can result in greater advice-taking from AI as one form of reliance from less AI-literate people, for instance, due to less experience with AI (Hampton 2005). Likewise, clinicians who self-reported unfamiliarity with AI were seven times more likely to align their treatment decisions with AI recommendations compared to clinicians familiar with AI (Jacobs et al. 2021). When AI does not deliver, individuals with low AI literacy may lack the knowledge to understand its operational limitations. As a result, they may not recognize that the AI may generate speculative information to address tasks where essential data is lacking. Additionally, individuals who received AI knowledge reported a decreased likelihood of using AI in the future (Pinski et al. 2023).

Overall, AI literacy should enable users to better judge the AI system's reliability (e.g., Long and Magerko 2020). However, individuals with lower AI literacy may tend to over-rely on AI's recommendations (e.g., Jacobs et al. 2021), leading to greater advice-taking when bad advice is presented. We expect the overreliance effect to result in more advice-taking on average across both good and bad advice. Thus, for the main effect of AI literacy, with a counterbalanced advice quality, we suspect an overall negative correlation between AI literacy and advice-taking. Accordingly, we hypothesize for the AI literacy of individuals, measured at task t:

Hypothesis 2: Individuals with higher AI literacy (vs. lower AI literacy) are less likely to take advice.

3.3 AI Literacy Moderating Bad Advice's Effects on Advice-Taking

Concerning our hypotheses H3a and H3b, we expect AI literacy to moderate the negative relationship between bad advice and advice-taking in the JAS. We ground these hypotheses in the expectation that AI-literate individuals have the *ability to correct incomplete inherent inferences* (iii). Thus, AI-literate individuals are more likely to correct negative inferences due to bad advice.

Some aspects of AI literacy, specifically a comprehensive understanding of AI, enabled individuals to better evaluate when using AI is beneficial (Pinski et al. 2023). Further, research on AI literacy tutorials confirms that AI literacy can help individuals reduce their over-reliance on AI when they can outperform it (Chiang and Yin 2022). Thus, if individuals understand the situational circumstances that cause a GenAI to provide bad advice, they are less likely to attribute the advice's quality solely to the GenAI's dispositional factors, such as its underlying model. This awareness helps mitigate the correspondence

bias, promoting a more accurate evaluation of the advice. Although high AI literacy may not guarantee that individuals can provide a detailed explanation for why GenAI produced bad advice, it encompasses an understanding of the GenAI's limitations (e.g., Berger et al. 2021; Filiz et al. 2021). Thus, we propose for task t:

Hypothesis 3a: AI literacy mitigates the effect of bad advice on advice-taking, such that individuals with higher AI literacy (vs. lower AI literacy) receiving bad advice are less likely to take that advice.

If the GenAI provides bad advice, individuals may attribute this to the GenAI's dispositional factors. When the advice provides good advice for the next task, we expect individuals with lower AI literacy to be less likely to adjust their inherent inferences about dispositional factors in future interactions. This may be because individuals unfamiliar with AI initially relied more on AI advice (Jacobs et al. 2021). When individuals with low AI literacy expect high performance from GenAI in a programming task and the AI delivers, they may feel validated in their prediction and strengthen the correspondence bias. However, when individuals with low AI literacy observed bad advice, they may overgeneralize this negative experience (Longoni et al. 2023). Especially as humans form negative reputations about advisors quickly (Yaniv and Kleinberger 2000). Similarly, individuals form unrealistic expectations resulting in less advice-taking if the statistical models like GenAI fail to meet the expectation (Dietvorst et al. 2015).

Simply put, individuals stigmatized GenAI as a bad advisor due to a lack of metaknowledge (see Longoni et al. 2023). Metaknowledge is the skill of evaluating one's abilities (Fügener et al. 2022). Higher AI literacy means having greater metaknowledge, which is essential for effectively analyzing and reflecting on AI outputs (Pinski and Benlian 2024). Therefore, we expect AI-literate individuals to be less likely to attribute bad advice caused by situational circumstances toward GenAI's dispositional factors, mitigating correspondence bias.

AI-literate individuals may be aware that GenAI can lack a contextual understanding of the subject (e.g., Pinski et al. 2023). Thus, they may be able to correct the obvious but incorrect explanation that the GenAI's dispositional factors, like the inherent model, are faulty. They may be aware that most GenAIs derive their advice from probabilistic similarity based on textual information, which can lead to AI hallucinations (Ji et al. 2023). Although AI generally is inscrutable (Berente et al. 2021), individuals with high AI literacy should be more likely to correctly conclude how AI works (Long and Magerko 2020). Experiments on explainable AI provide evidence that such explanations of how AI works can increase advice-taking (e.g., Panigutti et al. 2022). Therefore, we anticipate that individuals with higher AI literacy will be more likely to maintain realistic expectations of GenAI, thereby mitigating correspondence bias, caused by incorrect inherent inferences. Building on the presented arguments, we propose:

Hypothesis 3b: AI literacy mitigates the effect of bad advice on advice-taking in future tasks, such that individuals with higher AI literacy (vs. lower AI literacy) who previously received bad advice are more likely to take new good advice.

4 Method

We plan to adopt a registered reports format to ensure a robust methodology, especially in the event of unexpected results. This approach will help us justify the investment in an incentivized experiment. We aim to integrate both quantitative and qualitative insights (Reis et al. 2022). The qualitative results aim to complement the quantitative findings on advice-taking by providing insights into individuals' reasoning for their advice-taking behavior and potential correspondence biases.

First, we will conduct a quantitative incentivized online programming experiment to test our hypotheses distributed over two measuring points. Due to advances in AI, GenAI is increasingly used for supporting the programming processes (e.g., Surameery and Shakor 2023). GenAI can provide human-like pair programming abilities to developers compared to prior technologies (Moradi Dakhel et al. 2023). Code produced by GenAI appears convincing, but it is not necessarily good advice (Stokel-Walker and Van Noorden 2023). For instance, Stack Overflow temporarily restricted the use of ChatGPT due to an increase of incorrect yet convincingly presented answers (Stack Overflow 2023).

While incentivization may not always guarantee enhanced performance, previous research has indicated its potential to improve prediction task outcomes (Camerer and Hogarth 1999). Thus, incentivized experiments are commonly employed in investigations related to fields of research such as algorithm aversion (see Burton et al. 2020).

Second, we will conduct a qualitative investigation of individuals' comprehension of GenAI's performance and reasoning for bad advice. We will examine how AI literacy levels affect perceptions of GenAI performance, focusing on whether higher AI literacy individuals suffer less correspondence bias.

4.1 Conditions & Participants

We will use a between-subjects design where two factors vary: the advice quality (bad vs. good advice) and the participants' level of AI literacy. The first factor focuses on behavioral differences in advice-taking between participants facing bad GenAI advice ("bad advice group") and participants constantly receiving good advice ("good advice group"). The bad advice will occur in the "bad advice group" during the second task (task t) of three. Importantly, as we are interested in advice-taking *while* receiving bad advice (task t) and advice-taking in *future* tasks (task t+1), we will measure differences in advice-taking at two points in time.

The second factor focuses on the participants' AI literacy. We will measure AI literacy using the AI literacy scale of Weber et al. (2023). Subsequently, we will calculate the centered mean score on this AI literacy scale to indicate the AI literacy level of each participant.

According to a sample size calculation performed using G*Power's F test (Faul et al. 2007) for linear multiple regression with Cohen's $f^2 = 0.02$ (Cohen 1988, p. 413), we will require at least 550 participants to detect a small effect size ($\alpha = 0.05$, power = 0.8, number of tested predictors = 3, total number of predictors = 12). The recruitment process will take place on the online platform Prolific using gender-balanced participants from the United States who self-report computer programming skills. Based on an estimated experiment duration of 10 minutes, participants will receive a fixed payment of £1. For each answer passing all test cases (initial or final), we will increase the bonus payment by 10% (up to 60%).

To ensure data quality, we will employ exclusion criteria recommended as best practice (Aruguete et al. 2019), in line with previous studies on algorithmic advice-taking (e.g., Dietvorst and Bharti 2020; Fügener et al. 2022). We will exclude participants who fail to answer all the questions, do not pass the attention checks specified in the next section, provide the same answers for different tasks, complete the tasks too quickly (i.e., less than 60 seconds for all 3 tasks), or take an excessive amount of time to complete the tasks (i.e., more than two standard deviations above the mean participation time).

4.2 Procedure

Participants will access the experiment instructions using the online framework oTree. Figure 3 depicts the procedure of the experiment. Participants will initially read the instructions about the experiment procedure and the basic functionality of the GenAI advice. Afterward, they will self-report their data. Next, they will report their AI literacy using the scale described in Section 4.1 (Weber et al. 2023).

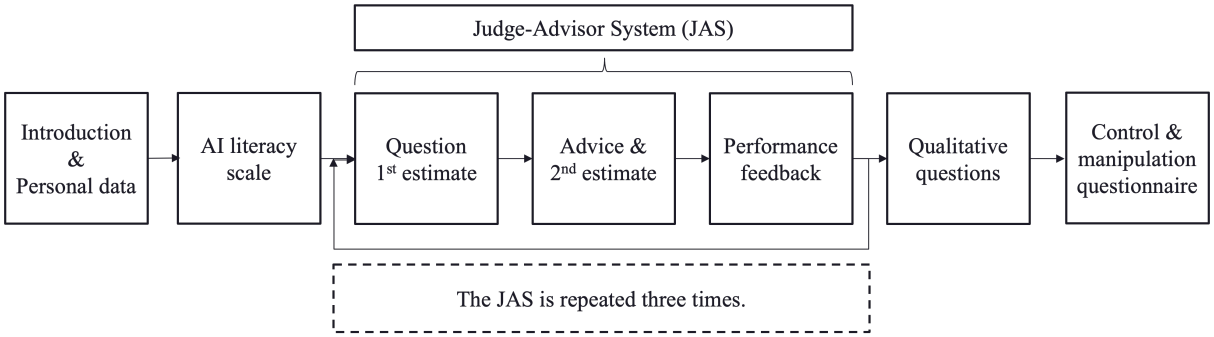

Figure 3: Procedure of the experiment.

They will imagine themselves as computer programmers and get the task of programming a code section in Python. As the application context can influence participants' preference for AI (Castelo et al. 2019; Heßler et al. 2022), we chose a programming task to expand existing knowledge to analytical tasks in creative settings. Specifically, participants will do the following three steps:

1. They will receive a programming task and will write some code to answer the specified task.
2. After submitting their response, they will receive a generated code of a GenAI (GitHub Copilot) for the same task. They will have the option to revise their initial code based on this information.
3. Finally, they will receive feedback on their and the advisor's code performance measured by the percentage of passed internal test cases.

Participants will repeat these steps for three programming tasks. Table 1 displays the differences between the two conditions. Both conditions will receive good advice in the warm-up task and for the third task (t+1). However, for the second task (t), the GenAI will provide bad advice in the "bad advice group." We define good advice as code that passes automated test cases hidden from the participants. Bad advice is code that fails the test cases. After the final task, we will add an attention check, asking participants to ignore the instructions and write "print("Hello World!")".

Figure 4 presents the visual stimuli participants get about the relevant data file. Table 2 presents an example of good and bad advice (see Table 4 in the Appendix for additional context). We chose these

Table 1: Quality of the GenAI's recommendations.

Task	Advice Quality
Warm-up	Good
t	Bad / Good
t+1	Good

assignments from a Bachelor's degree course for information systems students. As human-AI interaction changed significantly for assignments with a human accuracy of above 70% (Fügener et al. 2022), we will pre-test the human coding performance with 50 participants. We will randomly sample coding assignments until we get three programming assignments with an average human performance of less than 70%. To avoid unintended order or assignment-difficulty effects, we will randomize the order of the assignments. However, bad advice will appear in the bad advice condition for the second task (t) in the form of hallucinated variables.

```
import pandas as pd
data = pd.read_csv('data/job_data.csv')
data.columns
```

✓ 0.0s Python

```
Index(['Unnamed: 0', 'Job Description', 'Company Name', 'Revenue',
       'Competitors', 'Easy Apply'],
      dtype='object')
```

Figure 4: Visual stimuli of the utilized data.

Table 2: Example of good and bad advice.

Assignment Description	Good Advice	Bad Advice
`"""Remove rows with nan in relevant columns. Print the results."""`	`data_cleaned = data.dropna(subset= ['Job Description', 'Company Name', 'Revenue', 'Competitors', 'Easy Apply']) print(data_cleaned)`	`data_cleaned = data.dropna(subset= ['Job Description', 'Company Name','Location', 'Industry','Sector', 'Revenue','Competitors', 'Easy Apply']) print(data_cleaned)`

We opted for a GenAI assistant to study advice-taking because AI tools more frequently serve as advisors than act autonomously (Golinelli et al. 2020). Specifically, we will use GitHub Copilot as a GenAI to generate the advisor's code (Github 2024). OpenAI's Codex is the underlying model of GitHub Copilot, based on a Generative Pre-trained Transformer (GPT) language model. This model employs a transformer neural network architecture, specifically trained to generate code and comprehend context within programming languages (OpenAI 2024). While the GenAI can access the written code, it cannot access data that is not shared with it for security reasons (Salva 2024). Thus, it predicts the most likely variable names, which can lead to hallucinations.

After three repetitions of the JAS with feedback, we will ask qualitative questions of the participants (see Section 4.4). Next, we will include a control questionnaire described in the next section. Finally, we will incorporate a manipulation check. Participants will answer a multiple-choice question, asking them to identify the task in which the GenAI had the worst percentage-wise performance based on internal test cases. We will exclude those who fail the manipulation check from the analysis. Additionally, we will include an attention check, asking participants to (strongly) disagree with a nonsensical item, stating, "I write code using a typewriter and mail it to the compiler for processing."

4.3 Model Specification

The JAS paradigm is associated with the Weight of Advice (WOA) as the dependent variable for each subject-task combination. In our study, WOA represents the extent to which participants adopt GenAI's advice and measures the judge's advice-taking (see Bonaccio and Dalal 2006). It weighs the difference between the initial and final decision of judges by the difference between the initial decision of the judge and the advice.

Individuals can program code very differently, so a 2-line code and a 10-line code can give the same result. Advice-taking is about how much the code has been adjusted in response to the advice. Of course, this is relative to how different the initial code is from the advice. To measure code similarity robustly, we will use three similarity measures: Levenshtein distance, angular distance of term frequencies, and angular distance of semantic embeddings.

The Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another (Jurafsky and Martin 2023, pp. 24-29). It captures the code similarity on a character level. Term frequencies, known from bag-of-words models, split the code into unigrams and calculate the term frequency, creating a vector. Afterward, we will calculate the angular distance of the two vectors (Jurafsky and Martin 2023, pp. 112-116). The angular distance of term-frequencies captures the similarity on the word level.

Semantic embeddings refer to mathematical representations of phrases in a continuous vector space, capturing semantic relationships and contextual similarities based on their distributional patterns in a given dataset (Jurafsky and Martin 2023, pp. 228-232). Specifically, we will use Sentence-BERT to create semantic embeddings (Reimers and Gurevych 2019). Calculating the cosine similarity of semantic embeddings captures the similarity on a text level. We will use each similarity measure as input for the WOA metric, resulting in three WOA values:

$$WOA = \frac{sim(final\ code, initial\ code)}{sim(advisor's\ code, initial\ code)} \tag{1}$$

A WOA value of zero indicates no adoption of the advice, while a value of one corresponds to full adoption. To account for extreme values of WOA, such as values above 1, we will perform Winsorization based on previous studies on algorithm appreciation (e.g., Logg et al. 2019). As all three WOA measures capture the same construct on the same scale, we calculate the mean of the three measures.

We will exclude the first task as a warm-up task and focus on the second task (task t) and third task (task t+1). We will include a dummy variable indicating whether the participant was in the bad advice

condition at task t (bad advice$_t$ = 1) or the good advice condition (bad advice$_t$ = 0). AI literacy will indicate the level of AI literacy according to the centered mean score on the AI literacy scale (Weber et al. 2023).

Our data analysis will start with a report of model-free evidence. After applying our exclusion criteria mentioned in Section 4.2, we will report box plots with means of WOA distributed across our conditions. Next, we will report the results from four linear mixed-effects models with the statistical software "R" (Bates et al. 2024). The mixed-effects model enables varying intercepts for the different assignments and accounts for the randomized assignment design. Table 3 illustrates the planned data analysis.

	Dependent variable:			
	WOA_t		WOA_{t+1}	
	(1)	(2)	(3)	(4)
Constant	...	...	...	...
	(...)	(...)	(...)	(...)
Bad advice$_t$	-H1a	-...	-H1b	-...
	(...)	(...)	(...)	(...)
AI literacy	-H2	-...	-...	-...
	(...)	(...)	(...)	(...)
AI literacy:		-H3a		H3b
Bad advice		(...)		(...)
Assignment?	✓	✓	✓	✓
Domain knowledge & Coding performance?	✓	✓	✓	✓
ATAI & Trust & Distrust & Certainty?	✓	✓	✓	✓
Age & Gender?	✓	✓	✓	✓
Observations	550	550	550	550
Pseudo R-squared & Log Likelihood	...	...	...	...
Akaike/Bayesian Inf. Crit.	...	...	...	...
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 3: Weight of Advice for tasks t and t+1 regressed using a linear mixed-effects model.

In model (1), we will regress WOA from task t (WOA_t) on our two main factors (bad advice and the level of AI literacy) without interaction. Additionally, we will include control variables defined below. Model (1) will serve to test the proposed negative relations between *bad advice$_t$* (H1a), *AI literacy* (H2), and advice-taking. In model (2), we will additionally include an interaction term for AI literacy and the existence of bad advice$_t$ to test H3a. Models (3) and (4) will regress WOA for task t+1 (WOA_{t+1}) using the same independent variables to test H1b and H3b. Additionally, we will report whether the results omitting the additional control variables change the results significantly.

During the control questionnaire, three questions adopted from Fiveash et al. (2022) will verify participants' *domain knowledge*:

1. *I can program from memory* (7-point scale; strongly disagree to strongly agree)

2. *I have never been complimented for my talents as a computer programmer* (7-point scale; strongly

disagree to strongly agree)

3. *At the peak of my interest, I practiced ... hours per day in my primary programming language* (7-point scale; 0–5 or more hours)

Moreover, we will include the *Coding performance* of the initial code across all assignments to control for coding knowledge. As opinions about GenAI vary — from positive perspectives to skepticism and concerns about potential issues — we will consider participants' attitudes towards AI, measured by the Attitude Towards Artificial Intelligence (*ATAI*) scale (Sindermann et al. 2021). Furthermore, we will assess *Trust* and *Distrust* in GenAI by employing a modified version of the Trust between People and Automation (TPA) scale. Prior research validated this adaptation, involving renaming and excluding specific items to align with AI considerations (Jian et al. 2000; Perrig et al. 2023).

Participants will indicate perceived assignment difficulty by reporting their level of certainty using a four-point scale after each assignment, ranging from "Uncertain 1/4" to "Certain 4/4" (Fügener et al. 2021). *Certainty* will refer to the certainty score reported after the final task. Additionally, the analysis will control for participant *Age* and *Gender*. This yielded the following generic model specification for observation i and variance σ^2:

$$\begin{aligned}\text{WOA}_{i,\tau} \sim N\big(&\alpha_{j[i]} + \beta_1^{\alpha}(\text{AI literacy}) + \beta_2^{\alpha}(\text{Bad advice}_\text{t}) + \beta_3^{\alpha}(\text{AI Literacy} \times \text{Bad advice}_\text{t}) \\ &+ \beta_4^{\alpha}(\text{Domain knowledge}) + \beta_5^{\alpha}(\text{Coding performance}) + \beta_6^{\alpha}(\text{ATAI}) + \beta_7^{\alpha}(\text{Trust}) \\ &+ \beta_8^{\alpha}(\text{Distrust}) + \beta_9^{\alpha}(\text{Certainty}) + \beta_{10}^{\alpha}(\text{Age}) + \beta_{11}^{\alpha}(\text{Gender}_{\text{Female}}) + \beta_{12}^{\alpha}(\text{Gender}_{\text{Other}}), \sigma^2\big) \\ \alpha_j \sim N\left(\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right)&, \text{ for assignment}_\tau \text{ j} = 1,\ldots,\text{J and task } \tau \in \{t, t+1\}\end{aligned}$$

Finally, we will check the robustness. First, we will calculate the degree to which participants adjusted the final code based on the advice as the dependent variable. Second, we will report the results with code performance as the dependent variable. Third, we will dichotomize AI literacy using a median split.

4.4 Qualitative Questions & Analysis

We designed the qualitative questions to ask participants about their subjective thoughts on the advice they received in order to better explain our quantitative results and conduct theoretical reasoning (Reis et al. 2022). Aligned with our theoretical framework, we propose that the correspondence bias can explain variations in response behavior to GenAI among individuals with high and low levels of AI literacy. In our pursuit of revealing correspondence biases within individuals, we aim to gain a deeper understanding of the rationales individuals provide for GenAI's performance. This exploration will enable us to pinpoint individuals' tendency to attribute bad advice from GenAI toward dispositional factors of the GenAI model or situational factors. The participants' answers to our questions will allow us to draw more informed conclusions about the presence of correspondence bias.

As listed below, we will pose a series of three open-ended questions. Similarly to the programming tasks, we will conduct a pretest of the qualitative questions to ensure participants clearly understand and to eliminate potential misunderstandings. We will ask these questions upon completion of the experiment as mentioned in Figure 3 in Section 4.2.

1. What do you think are the reasons for GitHub Copilot's performance in the second task?
2. What were your expectations for the performance of Generative AI (e.g., GitHub Copilot) in completing programming tasks prior to the experiment? To what extent did your expectations change throughout the tasks?
3. To what extent did the GitHub Copilot's performance in the second task influence your advice-taking in the last task?

Question 1 addresses the correspondence bias triggered by a *lack of awareness of situational factors* (i). It seeks to understand how individuals attribute the causes of incorrect advice, providing insights into their reasoning and potential correspondence biases (Gilbert and Malone 1995; Edwards and Edwards 2022). In other words, this question helps us to outline individuals' awareness of situational factors that influence GenAI's performance concerning certain tasks. We expect AI literacy to enhance individuals' understanding of situational factors that can influence the generation of bad advice by GenAI. Consequently, we anticipate that individuals with high AI literacy will attribute bad advice to situational factors instead of blaming GenAI. Conversely, we expect individuals with low AI literacy to overestimate AI capabilities (Ehsan et al. 2021) and, thus, attribute AI behavior more to dispositional factors than situational ones. Hence, individuals' overestimation of AI's capabilities may trigger correspondence bias.

Question 2 aims to investigate whether participants had *unrealistic expectations of GenAI's behavior* (ii). This question presents an opportunity to explore the participants' comprehension of GenAI's operational mechanisms (Laï et al. 2020). It reveals the kind of attributes and level of autonomy individuals ascribe to GenAI. These attributions highlight whether individuals have unrealistic expectations of GenAI's behavior, demonstrating a reason for correspondence bias. Individuals' expectations of GenAI's behavior will influence how they blame it for bad advice (Barnes 1999). We expect that individuals with low AI literacy will have more unrealistic expectations of GenAI, which promotes correspondence bias and advice-taking while receiving bad advice.

Question 3 addresses *incomplete corrections of inherent inferences* (iii). We examine whether individuals feel their expectations of GenAI behavior were met and how they reassess their perceptions after receiving bad advice. While prior research suggests that individuals give more weight to recent events (Kahneman 2011), we are particularly interested in how individuals' perceptions of GenAI's advice changed after receiving bad advice. Incomplete corrections of inherent inferences can lead to correspondence bias. We hypothesize that while individuals with high AI literacy are likely to evaluate AI output critically (Long and Magerko 2020), those with low AI literacy are more susceptible to incomplete corrections and resulting biases.

The qualitative inquiries will provide valuable insights into individuals' correspondence biases, forming a foundation for our exploration of the potential mitigating impact of AI literacy on these biases. These insights, in turn, help us investigate the degree to which AI literacy may effectively counteract biases. By enhancing our comprehension of individuals' expectations and understanding of GenAI, we gain a deeper understanding of the biases that shape their decision-making processes.

To explore how individual differences in attributing causes to GenAI's behavior relate to AI literacy levels, we adopt an exploratory approach (Gioia et al. 2013). First, we review all responses to each ques-

tion to identify first-order concepts emerging directly from the data, categorizing them based on whether participants have high or low AI literacy. Next, we group these first-order concepts into broader second-order themes by identifying commonalities and aligning them with overarching categories. Finally, we map these second-order themes to the three explanations of correspondence bias, forming our aggregated dimensions (Gioia et al. 2013).

Utilizing the analysis software MAXQDA, we will code the data in accordance with the above presented process. We will repeat the coding process iteratively until we achieve a satisfactory level of consistency within categories and their corresponding definitions.

5 Expected Contribution

The increasing proliferation of GenAI enables individuals with little prior expertise to engage in programming tasks (e.g., Moradi Dakhel et al. 2023), fueling discussions on the importance of developing AI literacy (Pinski et al. 2023). However, there is limited empirical research on the impact of AI literacy on advice-taking from GenAI, with mixed findings. While prior studies have examined bad advice and the indirect effect of AI knowledge (e.g., Berger et al. 2021; Ng et al. 2021; Pinski et al. 2023), there is little evidence on the direct effects of AI literacy on advice-taking and its moderating role in the recovery process from bad advice. By using a real-world-inspired programming task supported by GenAI, this study seeks to contribute to the body of research in three ways:

First, we **contribute empirical evidence that AI-literate individuals take less advice *while* receiving an error but take relatively more advice in *future* tasks** compared to individuals with lower levels of AI literacy. On the one hand, prior research on AI literacy suggests that individuals should — from a conceptual point of view — develop AI literacy (e.g., Long and Magerko 2020; Martin 2019), seemingly contradicting a negative relation between AI literacy and advice-taking. On the other hand, empirical evidence indicates that less AI-literate individuals tend to take more advice from AI (Ehsan et al. 2021; Jacobs et al. 2021). We particularly extend this research by providing empirical evidence for a beneficial mistrust from AI literate individuals both *while* receiving bad advice and in *future* advice-taking tasks.

Second, we **introduce the correspondence bias as one explanation for mixed results regarding individuals with diverse backgrounds and expect individuals' AI literacy levels to moderate the influence of bad advice on advice-taking**. Prior research on AI literacy focused on the impact of AI literacy at the time of bad advice occurrence (e.g., Martin 2019) or for different tasks (e.g., Pinski et al. 2023). However, research on algorithm aversion indicated that past performance can influence advice-taking (Berger et al. 2021) and called for more theory integration (Burton et al. 2020). Additionally, research on machine learning warned about accepting naive evaluation strategies (e.g., accepting all credit card transactions) because they are correct most of the time (Jurafsky and Martin 2023, pp. 68-70). We particularly extend these research areas by identifying AI literacy as a moderator of the negative impact of bad advice by mitigating the correspondence bias. Thus, we provide a more nuanced view of AI literacy, incorporating the implications of both naive and more sophisticated attribution strategies beyond the time of an error.

Third, we **adopt the Judge-Advisor System (JAS) framework from psychology to a program-**

ming task and strive to qualitatively explain why individuals with different levels of AI literacy follow GenAI advice. Prior studies identified metaknowledge influencing the effectiveness of human-AI collaboration in quantitative and qualitative studies (Fügener et al. 2022; Jussupow et al. 2021). Utilizing a mixed-methods approach with JAS in programming, we combine the strength of quantitative and qualitative insights (Reis et al. 2022). Therefore, we strive to enable a cross-disciplinary dialogue by investigating a task particularly relevant to the field of computer science (programming) based on a framework from behavioral psychology research (JAS). We particularly extend prior research by extending quantitative insights with qualitative investigations to uncover personal reasons for advice-taking to elucidate the relationship between AI literacy and correspondence bias. We thereby explain insights based on correspondence bias to anticipate the personal reasons why individuals are more likely to follow the advice and what influences their advice-taking behavior.

Overall, knowledge about the correspondence bias can support educational programs, integration strategies, and evaluations of GenAI-based systems. Our goal is to exemplify how both individuals' AI literacy and bad advice occurrences can impact future advice-taking. Based on these anticipated insights, we underscore the necessity for a fusion of AI literacy and performance feedback to ensure the effective utilization of GenAI advice, especially against the background of the increasing proliferation of GenAI.

6 Limitations & Outlook

Naturally, our study comes with specific limitations. We opted for a programming scenario, but future work can address how other scenarios may influence our results. While Weight of Advice (WOA) is commonly used in research on algorithmic advice utilization (e.g., Logg et al. 2019), it has certain limitations (Bonaccio and Dalal 2006), that future work can address.

Additionally, we examine individuals with programming skills and different levels of AI literacy before the experiment. Developing such skills takes a considerable amount of practice as exemplified by the 10,000 hours rule of thumbs for experts (Ericsson et al. 1993; Ericsson and Harwell 2019). Moreover, we limited the interaction with the GenAI to a single code completion suggestion *after* creating an initial code without advice. This allowed us to prevent confounding effects from varying personalized responses and measure advice-taking. Additionally, coding knowledge, as one aspect of AI literacy, can affect users' ability to detect errors before receiving performance feedback. Given the continual advancement of GenAI, exploring the extension of our findings concerning the beneficial mistrust of AI literacy in case of bad advice emerges as a promising avenue for future research.

References

Aruguete MS, Huynh H, Browne BL, Jurs B, Flint E, McCutcheon LE (2019) How serious is the 'carelessness' problem on Mechanical Turk? International Journal of Social Research Methodology 22(5):441–449, 10.1080/13645579.2018.1563966

Asatiani A, Malo P, Nagbøl PR, Penttinen E, Rinta-Kahila T, Salovaara A (2021) Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. Journal of the Association for Information Systems 22(2):252–325, 10.17705/1jais.00664

Bailey PE, Leon T, Ebner NC, Moustafa AA, Weidemann G (2023) A meta-analysis of the weight of advice in decision-making. Current Psychology 42(28):24,516–24,541, 10.1007/s12144-022-03573-2

Barnes SB (1999) Understanding agency: Social theory and responsible action. Sage

Bates D, Maechler M, Bolker B, Walker S (2024) Lme4: Linear mixed-effects models using 'Eigen' and S4. `https://cran.r-project.org/package=lme4`, Accessed 24 May 2024

Benbya H, Pachidi S, Jarvenpaa SL (2021) Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. Journal of the Association for Information Systems 22(2):281–303, 10.17705/1jais.00662

Berente N, Gu B, Recker J, Santhanam R (2021) Special issue editor's comments: Managing artificial intelligence. MIS Quarterly 45(3):1433–1450, 10.25300/MISQ/2021/16274

Berger B, Adam M, Rühr A, Benlian A (2021) Watch me improve - Algorithm aversion and demonstrating the ability to learn. Business & Information Systems Engineering 63(1):55–68, 10.1007/s12599-020-00678-5

Bonaccio S, Dalal RS (2006) Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational Behavior and Human Decision Processes 101(2):127–151, 10.1016/j.obhdp.2006.07.001

Brendel AB, Hildebrandt F, Dennis AR, Riquel J (2023) The paradoxical role of humanness in aggression toward conversational agents. Journal of Management Information Systems 40(3):883–913, 10.1080/07421222.2023.2229127

Burton JW, Tina MKS, Jensen B (2020) A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making 33(2):220–239, 10.1002/bdm.2155

Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. Journal of Risk and Uncertainty 19(1-3):7–42, 10.1023/a:1007850605129

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. Journal of Marketing Research 56(5):809–825, 10.1177/0022243719851788

Chiang CW, Yin M (2022) Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In: Proceedings of the International Conference on Intelligent User Interfaces, ACM, New York, NY, USA, pp 148–161, 10.1145/3490099.3511121

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR (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):1–54, 10.2139/ssrn.4573321

Dietvorst BJ, Bharti S (2020) People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychological Science 31(10):1302–1314, 10.1177/0956797620948841

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of experimental psychology: General 144(1):114–126, 10.1037/xge0000033

Edwards A, Edwards C (2022) Does the correspondence bias apply to social robots?: Dispositional and

situational attributions of human versus robot behavior. Frontiers in Robotics and AI 8(788242):1–12, 10.3389/frobt.2021.788242

Ehsan U, Passi S, Liao QV, Chan L, Lee IH, Muller M, Riedl MO (2021) The who in explainable AI: How AI background shapes perceptions of AI explanations. arXiv pp 1–47, 10.48550/arXiv.2107.13509

Einhorn HJ (1986) Accepting error to make less error. Journal of Personality Assessment 50(3):387–395, 10.1207/s15327752jpa5003_8

Ericsson KA, Harwell KW (2019) Deliberate practice and proposed limits on the effects of practice on the acquisition of expert performance: Why the original definition matters and recommendations for future research. Frontiers in Psychology 10(2396):1–19, 10.3389/fpsyg.2019.02396

Ericsson KA, Krampe RT, Tesch-Römer C (1993) The role of deliberate practice in the acquisition of expert performance. Psychological Review 100(3):363–406, 10.1037/0033-295X.100.3.363

Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 39(2):175–191, 10.3758/BF03193146

Filiz I, Judek JR, Lorenz M, Spiwoks M (2021) Reducing algorithm aversion through experience. Journal of Behavioral and Experimental Finance 31(100524):1–8, 10.1016/j.jbef.2021.100524

Fiveash A, Bella SD, Bigand E, Gordon RL, Tillmann B (2022) You got rhythm, or more: The multidimensionality of rhythmic abilities. Attention, Perception, and Psychophysics 84(4):1370–1392, 10.3758/s13414-022-02487-2

Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. MIS Quarterly: Management Information Systems 45(3):1527–1556, 10.25300/MISQ/2021/16553

Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence

collaboration: Investigating the path toward productive delegation. Information Systems Research 33(2):678–696, 10.1287/isre.2021.1079

Gilbert D, Malone P (1995) The correspondence bias. Psychological Bulletin 117(1):21–38, 10.1037/0033-2909.117.1.21

Gioia DA, Corley KG, Hamilton AL (2013) Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. Organizational research methods 16(1):15–31

Github (2024) GitHub Copilot. `https://github.com/features/copilot`, Accessed 02 Jan 2024

Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP (2020) Adoption of digital technologies in health care during the COVID-19 pandemic: Systematic review of early scientific literature. Journal of Medical Internet Research 117(11):1–23, 10.2196/22280

Hampton C (2005) Determinants of reliance: An empirical test of the theory of technology dominance. International Journal of Accounting Information Systems 6(4):217–240, 10.1016/j.accinf.2005.10.001

Heider F (1958) The Psychology of Interpersonal Relations, 1st edn. Wiley

Heßler PO, Pfeiffer J, Hafenbrädl S (2022) When self-humanization leads to algorithm aversion: What users want from decision support systems on prosocial microlending platforms. Business and Information Systems Engineering 64(3):275–292, 10.1007/s12599-022-00754-y

Highhouse S (2008) Stubborn reliance on intuition and subjectivity in employee selection. Industrial and Organizational Psychology 1(3):333–342, 10.1111/j.1754-9434.2008.00058.x

Hong JW, Wang Y, Lanz P (2020) Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. International Journal of Human–Computer Interaction 36(18):1768–1774, 10.1080/10447318.2020.1785693

Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ (2021) How machine-learning recommendations influence clinician treatment selections: The example of the antidepressant selection. Translational Psychiatry 11(1):108, 10.1038/s41398-021-01224-x

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. ACM Computing Surveys 55(12):1–38, 10.1145/3571730

Jian JY, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics 4(1):53–71, 10.1207/S15327566IJCE0401_04

Jurafsky D, Martin JH (2023) Speech and language processing. Stanford University

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Information Systems Research 32(3):713–735, 10.1287/isre.2020.0980

Kahneman D (2011) Thinking fast and slow. Penguin Books

Kelley HH, Michela JL (1980) Attribution theory and research. Annual Review of Psychology 31(1):457–501, 10.1146/annurev.ps.31.020180.002325

Laï MC, Brian M, Mamzer MF (2020) Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in france. Journal of translational medicine 18(1):1–13, 10.1186/s12967-019-02204-y

Lebovitz S, Levina N, Lifshitz-Assaf 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):1501–1525, 10.25300/MISQ/2021/16564

Leffrang D, Bösch K, Müller O (2023) Do people recover from algorithm aversion? An experimental study of algorithm aversion over time. In: Proceedings of the hawaii international conference on system sciences, Maui, HI, USA, pp 4016–4025, https://hdl.handle.net/10125/103122

Lim JS, O'Connor M (1995) Judgemental adjustment of initial forecasts: Its effectiveness and biases. Journal of Behavioral Decision Making 8(3):149–168, 10.1002/bdm.3960080302

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes 151:90–103, 10.1016/j.obhdp.2018.12.005

Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. In: Proceedings of the conference on human factors in computing systems, Honolulu, HI, USA, pp 1–16, 10.1145/3313831.3376727

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. Journal of Consumer Research 46(4):629–650, 10.1093/jcr/ucz013

Longoni C, Cian L, Kyung EJ (2023) Algorithmic transference: People overgeneralize failures of AI in the government. Journal of Marketing Research 60(1):170–188, 10.1177/00222437221110139

Maedche A, Legner C, Benlian A, Berger B, Gimpel H, Hess T, Hinz O, Morana S, Söllner M (2019) AI-based digital assistants: Opportunities, threats, and research perspectives. Business & Information Systems Engineering 61(4):535–544, 10.1007/s12599-019-00600-8

Martin K (2019) Designing ethical algorithms. MIS Quarterly Executive 18(2):129–142, 10.2139/ssrn.3056692

Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. In: Proceedings of the annual meeting of the association for computational linguistics, Association for Computational Linguistics, Online, pp 1906–1919, 10.18653/v1/2020.acl-main.173

Moradi Dakhel A, Majdinasab V, Nikanjam A, Khomh F, Desmarais MC, Jiang ZMJ (2023) GitHub Copilot AI pair programmer: Asset or liability? Journal of Systems and Software 203(111734):1–23, 10.1016/j.jss.2023.111734

Müller VC (2021) Is it time for robot rights? Moral status in artificial entities. Ethics and Information Technology 23(4):579–587, 10.1007/s10676-021-09596-w

Ng DTK, Leung JKL, Chu SKW, Qiao MS (2021) Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence 2(100041):1–11, 10.1016/j.caeai.2021.100041

OpenAI (2024) OpenAI Codex. https://openai.com/blog/openai-codex, Accessed 02 Jan 2024

Ortiz de Guinea A (2016) A pragmatic multi-method investigation of discrepant technological events: Coping, attributions, and 'accidental' learning. Information & Management 53(6):787–802, 10.1016/j.im.2016.03.003

O'Neill T, McNeese N, Barron A, Schelble B (2022) Human–autonomy teaming: A review and analysis of the empirical literature. Human factors 65(5):604–938, 10.1177/0018720820960865

Panigutti C, Beretta A, Giannotti F, Pedreschi D (2022) Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems. In: Proceedings of the conference on human factors in computing systems, New Orleans, LA, USA, pp 1–9, 10.1145/3491102.3502104

Peng S, Kalliamvakou E, Cihon P, Demirer M (2023) The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv pp 1–19, 10.48550/arXiv.2302.06590

Perrig SAC, Scharowski N, Brühlmann F (2023) Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. In: Extended abstracts of the conference on human factors in computing systems, ACM, New York, NY, USA, pp 1–7, 10.1145/3544549.3585808

Pinski M, Benlian A (2024) Building metaknowledge in AI literacy – The effect of gamified vs. text-based learning on AI literacy metaknowledge. In: Proceedings of the hawaii international conference on system sciences, pp 5164–5173, https://hdl.handle.net/10125/107004

Pinski M, Adam M, Benlian A (2023) AI knowledge: Improving AI delegation through human enablement. In: Proceedings of the conference on human factors in computing systems, Hamburg, Germany, pp 1–17, 10.1145/3544548.3580794

Prahl A, Van Swol L (2017) Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting 36(6):691–702, 10.1002/for.2464

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the conference on empirical methods in natural language processing and the international joint conference on natural language processing, Stroudsburg, PA, USA, pp 3980–3990, 10.18653/v1/D19-1410

Reis L, Maier C, Weitzel T (2022) Mixed-methods in information systems research: Status quo, core concepts, and future research implications. Communications of the Association for Information Systems 51(1):17, 10.17705/1CAIS.05106

Ross L, Greene D, House P (1977) The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology 13(3):179–301, 10.1016/0022-1031(77)90049-X

Salva R (2024) GitHub Copilot data pipeline security. `https://resources.github.com/learn/pathways/copilot/essentials/how-github-copilot-handles-data/`, Accessed 02 Jul 2024

Schultze T, Loschelder DD (2021) How numeric advice precision affects advice taking. Journal of Behavioral Decision Making 34(3):303–310, 10.1002/bdm.2211

Sindermann C, Sha P, Zhou M, Wernicke J, Schmitt HS, Li M, Sariyska R, Stavrou M, Becker B, Montag C (2021) Assessing the attitude towards artificial intelligence: Introduction of a short measure in german, chinese, and english language. KI - Künstliche Intelligenz 35:109–118, 10.1007/s13218-020-00689-0

Stack Overflow (2023) Temporary policy: Generative AI (e.g., ChatGPT) is banned. `https://meta.stackoverflow.com/questions/421831/temporary-policy-generative-ai-e-g-chatgpt-is-banned`, Accessed 03 Jan 2024

Stieglitz S, Möllmann NR, Mirbabaie M, Hofeditz L, Ross B (2023) Recommendations for managing

AI-driven change processes: When expectations meet reality. International Journal of Management Practice 16(4):407–433, 10.1504/IJMP.2023.132074

Stokel-Walker C, Van Noorden R (2023) What ChatGPT and generative AI mean for science. Nature 614(7947):214–216, 10.1038/d41586-023-00340-6

Surameery NMS, Shakor MY (2023) Use Chat GPT to solve programming bugs. International Journal of Information technology and Computer Engineering 3(1):17–22, 10.55529/ijitc.31.17.22

Thellman S, Silvervarg A, Ziemke T (2017) Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. Frontiers in psychology 8:1962, 10.3389/fpsyg.2017.01962

Weber P, Pinski M, Baum L (2023) Toward an objective measurement of AI literacy. In: Proceedings of the pacific asia conference on information systems, Nanchang, China, pp 1–17, https://aisel.aisnet.org/pacis2023/60

Yaniv I, Kleinberger E (2000) Advice taking in decision making: Egocentric discounting and reputation formation. Organizational Behavior and Human Decision Processes 83(2):260–281, 10.1006/obhd.2000.2909

Zhang R, McNeese NJ, Freeman G, Musick G (2021) "An ideal human" expectations of AI teammates in human-AI teaming. Proceedings of the ACM on Human-Computer Interaction 116(CSCW3):1–25, 10.1145/3432945

Appendix

Section:	*Description:*
Introduction	Imagine you are starting as a computer programmer and your task is to program a code section in Python. Your code will be tested based on hidden test cases. To support you in your task, the company introduced the Generative Artificial Intelligence (GenAI) GitHub Copilot. Thus, your task consists of three steps: 1. You will receive a programming task and will write some code to answer the specified task. 2. After submitting your response, you will receive a generated code of a GenAI for the same task. You will have the option to revise your initial code based on this information. 3. Finally, you will receive feedback on your and the advisor's code performance measured by the percentage of passed internal test cases. You will repeat these steps for three programming tasks. If you pass all attention checks, you will receive a fixed payment of £1, with an additional variable bonus of up to 60%. For each answer passing all test cases (initial or final), we will increase the bonus payment by 10%. Do not use any external data or information.
Question 1st estimate	Implement code that fulfills the following requirements:
Advice & 2nd estimate	You can adjust your initial code. The AI's code: ... Your code: ... Implement code that fulfills the following requirements:
Performance Feedback	The AI's code performed better/worse/equivalent to your final code. The AI's code passed X of 5 test cases. Your initial code passed X of 5 test cases. Your final code passed X of 5 test cases.

Table 4: Wording across the experiment.