To Use or Not to Use: Impatience and Overreliance When Using Generative AI Productivity Support Tools

Han Qiao
Autodesk Research
Toronto, Ontario, Canada
h.qiao@mail.utoronto.ca

Jo Vermeulen
Autodesk Research
Toronto, Ontario, Canada
jo.vermeulen@autodesk.com

George Fitzmaurice
Autodesk Research
Toronto, Ontario, Canada
george.fitzmaurice@autodesk.com

Justin Matejka
Autodesk Research
Toronto, Ontario, Canada
justin.matejka@autodesk.com

Abstract

Generative AI has the potential to assist people with completing various tasks, but increased productivity is not guaranteed due to challenges such as uncertainty in output quality and unclear processing time. Through an online crowdsourced experiment (N=508), leveraging a "paint by numbers" task to simulate properties of GenAI assistance, we explore how, and how well, users make decisions on whether to use or not use automation to maximize their productivity given varying waiting times and output quality. We observed gaps between user's actual choices and their optimal choices and characterized these gaps as the "gulf of impatience" and the "gulf of overreliance". We also distilled strategies that participants adopted when making their decisions. We discuss design considerations in supporting users to make more informed decisions when interacting with GenAI tools and make these tools more useful for improving users' task performance, productivity and satisfaction.

CCS Concepts

• **Human-centered computing** → **Laboratory experiments**; **Empirical studies in HCI**.

Keywords

generative AI, decision-making, productivity, reliance, AI, automation, controlled experiment

ACM Reference Format:
Han Qiao, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2025. To Use or Not to Use: Impatience and Overreliance When Using Generative AI Productivity Support Tools. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714103

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714103

1 Introduction

Recent advancements in Generative AI (GenAI) have led to the development of a wide variety of GenAI support tools that assist people in completing various tasks across different domains (including writing, design, manufacturing, video production, education and programming). GenAI support tools have proven to effectively assist people with divergent thinking, especially associated with ideation and prototyping in early stages of design projects [31, 58, 69], as well as in end-to-end task completion with specifications and requirements, associated with tasks such as programming [11, 39], manufacturing [13] and landscape rendering [20]. However, despite their capabilities, GenAI tools require time to generate responses, often produce uncertain outputs, and demand additional effort from users to make further edits and correct mistakes [13, 32, 53, 61]. Therefore, the use of GenAI support tools does not always result in guaranteed improvements in productivity [39, 53].

For users seeking ideation support, or for novices who are not able to complete tasks without the assistance of GenAI, productivity may not be the primary concern. However, for users who are able to complete tasks like writing, 3D modeling and programming on their own but are now offered GenAI assistance, the uncertainty around the tools' impact on productivity raises important questions about how individuals engage with these tools to reach their optimal task performance. Therefore, in this paper, we explore the following research questions:

- RQ1: Given the goal of maximizing productivity, how good are people at making optimal decisions about when to use GenAI?
- RQ2: How do people decide when to use or not use GenAI tools?
- RQ3: How can the design of these systems better help users understand when to use or not to use GenAI to enhance their productivity?

We investigate people's decision making and reliance on GenAI-like tools through an online crowdsourced experiment. We designed a controlled environment where we could study the effects of latency and error rate in a setting similar to what users might find when using GenAI. More specifically, participants were asked to complete a series of "paint by numbers" tasks (Figure 1), simulating the type of tasks that users could complete on their own, but with which GenAI could also offer assistance. To simulate a GenAI tool,

we offered participants an *Assisted Fill* tool, which varied in error rate and latency, requiring participants to fix any mistakes made by the tool. A total of 508 participants were asked to decide when to use the Assisted Fill tool under different error rate and latency conditions. Based on participants' performance of completing the tasks manually, completing the tasks with Assisted Fill, and their decisions on when to use Assisted Fill, we quantitatively evaluated how well participants made optimal decisions.

Our results show that even when provided with performance information about Assisted Fill, there were gaps between participants' optimal choices and their actual decisions when aiming to maximize productivity. We characterize these two types of decision-making gaps as the *gulf of impatience* and the *gulf of overreliance*. We locate the gaps across all conditions and find that people made optimal decisions when automation had either high performance (low latency + low error rate) or low performance (high latency + high error rate). The gulf of impatience occurred when automation's performance was moderate (moderate latency + moderate error rate). The gulf of overreliance occurred especially when the automation excelled in one area, such as low latency, but performed poorly in the other, such as a high error rate. We conclude by discussing the implications for designing human-AI interaction that supports better decision making for improving users' task performance, productivity and user satisfaction.

In summary, our work offers the following four contributions:

- A study methodology to run controlled experiments for systematically testing characteristics of GenAI interaction using a "paint by numbers" web app as a proxy for GenAI support tools and human-AI interaction.
- A characterization of decision-making gaps as the *gulf of impatience* and the *gulf of overreliance*, illustrating common patterns around when users decide to use or not to use GenAI tools across various latency and error rate conditions.
- A discussion of key strategies people rely on when making these decisions around when to use or not to use GenAI tools for enhancing productivity.
- Design considerations for designing future GenAI tools and workflows that support better decision making to improve productivity, performance and user satisfaction.

2 Related Work

2.1 Generative AI Support Tools

GenAI offers abundant system and product design opportunities for assisting people in performing various tasks. Research projects have explored the potential application of GenAI in a wide range of domains such as writing [12, 13], programming [62], 2D design [61], 3D design [33], UI design [9], video production [58] and research [11] etc. In various studies, these GenAI supported systems and workflows have been shown to augment people's abilities, by reducing mental task load [58], enhancing creativity [34], and assisting in producing higher quality results [9, 20].

In the above systems and workflows, GenAI is used to support various types and stages of work. For instance, some GenAI support tools come in at earlier stages to support brainstorming [69], ideation [58, 69] and early stage communication between designers [20]. In these scenarios, GenAI is not required to produce outputs that match perfectly with specific requirements, but to provide a variety of outputs to support divergent thinking for inspiration. However, in other applications of GenAI, users are able to complete the task on their own, but GenAI support tools can assist with a portion of the task so that users can then build upon the GenAI outputs or fix any errors to meet requirements. Prior research has explored this strategy in domains such as designing 3D structures [13], logo design [61], writing [12], and programming [2, 39, 53]. Users of a GenAI logo design system explicitly mentioned the importance of final editing after getting the outputs so that they could change the logos to match their expectations [61]. Similarly, as illustrated in research around GenAI programming assistants, GenAI is seen as not useful when users have to spend too much time fixing errors in generated code to achieve correct functionality, or even just making sense of or validating the generated code [2, 30, 46].

In our study, we focus on the second type of tasks, in which users have to complete a task with requirements that need to be met, and GenAI support tools are leveraged to help them carry out that task from start to finish. This is in contrast with tasks that users cannot complete on their own without GenAI, or scenarios where GenAI acts as a source of inspiration. For those tasks in which GenAI tools support users in task completion, enhancing *productivity* becomes a particularly critical lens to evaluate the effectiveness of GenAI.

2.2 Human-AI interaction and Productivity

Previous work studying human-AI interaction has revealed a wide range of potential positive impacts of incorporating GenAI in existing workflows. For instance, these include studies that showed how GenAI could: enhance human creativity [31, 34], reduce mental task load [11, 62], be incorporated in teamwork settings to enhance discussion facilitation and exchange of ideas [11, 20].

Productivity is one key area of focus when evaluating the usability and utility of GenAI support tools. When evaluating their GenAI system for supporting short video script writing, Wang et al. [58] emphasized that their system was able to reduce users' time and effort. Ziegler et al. [70] surveyed GenAI programming assistant users on their perceived productivity and found that developers felt more productive using GenAI support tools and that self-reported perceptions were correlated with suggestion acceptance rate. Randomized control trials were also conducted by different groups of researchers, showing that participants reduced their coding task completion time when using GenAI programming assistants [46, 57].

However, an increase in productivity is not always guaranteed [53]. Vaithilingam et al. [57] found no significant improvement in task completion time when using Github Copilot and other researchers have also pointed out usability issues with GenAI in programming assistance [30], design [26, 42] and manufacturing [13]. Simkute et al. [53] linked loss of productivity to Bainbridge's "ironies of automation" [1].

Recognizing the unique impact of human-AI interaction on productivity, HCI researchers have reflected on what makes human-AI interaction difficult to design for and what factors influence productivity. As mentioned in Section 2.1, our study focuses on a particular type of GenAI support tools for assisting users in completing tasks with specific requirements (such as programming or manufacturing products), rather than just assisting in the brainstorming stage.

Below, we synthesize previously explored key characteristics and challenges of working with those particular types of GenAI productivity support tools into three categories: *output uncertainty*, *output manipulation*, and *output latency*.

- **Output Uncertainty**: Yang et al. [64] identified that designers often find designing with AI difficult due to uncertainty about the AI's capabilities and the complexity of the AI's outputs. They analyzed two attributes of AI uncertainty: capability uncertainty, referring to uncertainty surrounding what the system can do and output uncertainty, which refers to the complexity of the outputs that the system might generate. Liang et al. [30] also echoed these findings and pointed out that an important usability issue with GenAI assistance is experiencing difficulties controlling the output, which leads to output uncertainty.
- **Output Manipulation**: Past research has analyzed how users interact with GenAI support tools and found that users spend significant time validating, manipulating and optimizing AI output. Gmeiner et al. [13] found that one primary challenge for CAD users when using AI design tools is manipulating AI outputs. Their study showed that when AI supports manufacturing design works, users usually need to edit and rework AI outputs that contain aesthetic flaws, surface bumps, holes, or slightly twisted geometry for 3D models. Xiao et al. [61] designed a system for generating Typographic Logos, and their usability evaluation similarly highlighted participants' needs for post-generation editing. In a different domain, Yan et al. [62] studied programming assistance (GitHub Copilot, OpenAI GPT-4, Amazon CodeWhisper, IntelliCode Compose, CodeT5+) and observed that programmers spend a significant amount of time examining the generated code and encounter challenges of understanding, revising and fixing any mistakes. Similarly, Liang et al [30] found that manipulating AI output includes steps of understanding output, evaluating output, and either modifying output or giving up on output. Their results showed that some programmers spent too much time modifying results or found the output distracting.
- **Output Latency**: Processing delays and waiting for output to be generated is another characteristic that designers need to take into consideration when designing GenAI support tools. Liu et al. [34] pointed out that processing delays are a key characteristic of GenAI tools, and argued for taking such delays into consideration for system design. Mozannar et al. [39] studied how programmers use GenAI assistance and developed a taxonomy of common programmer activities when interacting with Copilot. Out of the 12 activities, they found "waiting for suggestions" on average accounted for 4.2% of the total time of task completion and that programmers often have to wait for suggestions to show up due to either latency or Copilot not kicking in to provide a suggestion. Lin & Martelaro [31] identified waiting time to obtain results from a model as one of the key challenges of working with AI, making the work no longer a seamless process.

Therefore, in our study, we simulate a GenAI system through a "paint by numbers" task that incorporates these three characteristics of GenAI that we discussed above. We then leverage this task to investigate how and how well people make decisions around when to use or not use GenAI when trying to maximize their productivity given these interaction challenges.

2.3 Reliance on AI

A rich body of research in HCI, psychology and other fields have explored the topic of reliance on automation [7, 35, 43, 44]. Automation is defined as the execution by a machine agent of a function that was previously carried out by a human [43]. When humans inappropriately rely on automation, including misuse, disuse, overreliance, or underreliance, the human-machine collaboration fails to achieve the optimal outcomes or results in negative consequences.

With the advancement and adoption of machine learning (ML) and AI-based applications, researchers started looking into human reliance on AI, with much focus on AI-assisted decision making: how people accept AI suggestions and the consequences of accepting erroneous suggestions or rejecting correct suggestions [45]. Prior research studied factors that impact people's reliance on ML models, including stated accuracy [27], observed accuracy [66], the model's confidence associated with each recommendation [68], and people's confidence in their agreement/disagreement with the model [35]. Some studies found that explanations of an ML model lead to more appropriate reliance on the model [27, 48, 63], while others observed that reliance varies across decision-making tasks, types of explanations, and user characteristics [3, 6, 24, 51, 59, 68].

Reliance on GenAI support tools is quite different from reliance on AI-assisted decision making due to differences in the types of tasks supported, ways of interacting with AI's output, and consequences of accepting or rejecting AI's suggestions. GenAI recommendations are applied to more open-ended tasks that could be completed with a variety of pathways, whereas AI-assisted decision making is applied to tasks that can usually be reduced to binary decision making [35, 51, 68]. Therefore, we see much value in advancing our understanding of human reliance on automation in the context of GenAI. Roy et al. [50] explored the impact of accuracy and "controllability" (how easily an automated result can be manually modified) on decisions of using or not using automation, shedding light on people's preferences on automation with characteristics of GenAI. Our study aims to deepen our understanding of people's reliance on GenAI through the focal point of productivity. Since GenAI recommendations can no longer be evaluated on a simple dichotomy of correct versus incorrect, as often found in studies of ML-assisted decision making, we explore people's reliance on GenAI when they try to maximize their productivity. Therefore, overreliance, previously defined as users accepting incorrect AI recommendations [45], in the context of our paper becomes users spending more time using GenAI to assist with completing a task than it would take to complete the task themselves. Similarly, underreliance, previously defined as users rejecting correct AI recommendations, becomes users spending more time completing a task on their own when leveraging GenAI could have helped them complete the task faster. Our study expands our understanding of human-GenAI interaction (and of human-AI interaction at large), through revealing how people rely on GenAI support tools, and

how and how well they make decisions on when to use or not to use GenAI support tools for productivity support.

3 Methodology

We wanted to investigate how and how well people decide whether to use or not to use an automation support tool that simulates characteristics of GenAI when trying to maximize productivity. Rather than using a true GenAI system where the outputs could not be easily controlled, we developed our own *simulated* Generative AI environment where we could carefully control the latency and error rates. We designed a "paint by numbers" task that resembles the characteristics of the type of tasks that we focus on. As discussed earlier, we focus on tasks where GenAI is leveraged to support users in completing tasks with specific requirements, such as those required in manufacturing, 3D modeling and coding, rather than those that merely provide brainstorming ideas and inspirations. As for the simulated GenAI support tool, we specifically incorporated characteristics of output uncertainty, output latency and output manipulation into an *Assisted Fill* tool offered to participants. The experiment assigned 508 participants into conditions for interacting with Assisted Fill with various latency and output accuracy. In the following sections, we explain in detail the paint by numbers task, the simulated GenAI tool, and the experiment design.

3.1 Simulated Environment: Paint by Numbers

The task that users need to complete in our experiment is a "paint by numbers" task, which we developed as a web app in JavaScript with data stored and managed through a MongoDB backend. The idea is for the user to fill in the correct colors by pressing the associated number key for each cell and to replicate a pixelated image that is provided on the side. As shown in Figure 1, users start with a blank grid of cells with only colored numbers on the left and then work towards filling in colors to recreate the pixelated image shown on the right (in this case, a duck).

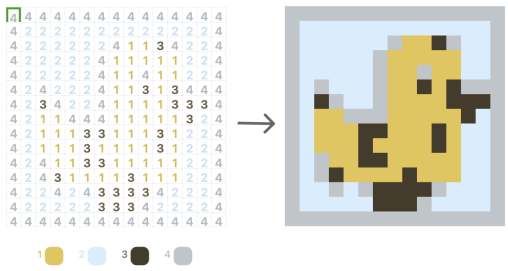

Figure 1: An example of an empty grid (left) and the final image presented to users (right) in a paint by numbers task.

As illustrated in Figure 2, users can use the arrow keys or mouse to move through grids of cells and fill in colors by pressing the corresponding number keys. Two modes of interaction are offered for completing the tasks. One is **Manual Fill**, in which users only rely on their keyboard and mouse to complete filling in colors for each cell. The other mode is **Assisted Fill**, in which GenAI is simulated. If Assisted Fill is chosen, users will wait for the colors to be automatically filled in and then correct any mistakes.

As mentioned before, we designed this task to simulate a human-AI interaction scenario in which GenAI is used to support users completing tasks with specific requirements. Users have to make sure that they fill in all of the cells correctly to complete the task. The Assisted Fill tool was designed to incorporate the three characteristics of GenAI tools that previous research highlighted as important challenges that impact productivity (see Section 2.2):

- **Output Uncertainty** is simulated by users not being able to predict patterns in terms of *where* there will be wrong colors in cells and *which* wrong color will be filled in as errors.
- **Output Manipulation** is simulated by requiring users to correct any mistakes made by the automation before completing the task. We also deliberately designed the fill color of the cells and the number text color to be subtly different (adjusting each RGB value by 0.5%), ensuring that mistakes are noticeable (due to mismatched text and background colors), but not overly obvious. This mimics how GenAI often introduces small errors that need to be detected and fixed by users, and may not be obvious to find.
- **Output Latency** is simulated by assigning users to various waiting times for Assisted Fill. The two modes of interaction, Manual Fill and Assisted Fill, are illustrated in Figure 3.

3.1.1 Design Decisions on Controllability vs. Realism. This study methodology using our "paint by numbers" interactive web app can act as an effective proxy to study GenAI support tools, being more practical to implement and providing more control compared to integrating a real GenAI model. Our methodology prioritized controllability and precision afforded by an (online) lab experiment, and we recognize that this approach comes at the expense of *realism* [38, p. 155]. McGrath argues that any methodology presents "both opportunities for gaining knowledge and limitations to that knowledge" [38, p. 154]. While researchers hope to gather evidence and produce insights that are generalizable, precise, and closely aligned with real-world contexts, we often face the inherent dilemma of not being able to achieve all dimensions simultaneously through one methodology. We believe that developing a method for a controlled experiment serves as a crucial step to understand human decision-making patterns during human-GenAI interaction. Once developed, this method could be reused, allowing researchers to replicate studies, collect data on a larger scale, understand differences caused by small changes in key factors of focus, and customize the design and methodology to accommodate other scenarios. For instance, we see this method potentially extending to other types of tools and tasks, including GenAI systems with varying levels of difficulty for detecting errors, UI designs that support different ways of presenting model information, or systems that afford users to multi-task.

3.1.2 Design Decisions on Error Detectability. We went through several rounds of iterative design among the authors and conducted pilot studies with experienced users of GenAI tools to make sure that our system resembles characteristics of GenAI. More specifically, we were trying to make sure the errors were recognizable but not overly obvious, requiring users to pay attention to identify errors while avoiding excessive time spent searching for them during the task. We tested out different combinations of cell colors and text colors so that the visual cues, differences in hues, and shades are not

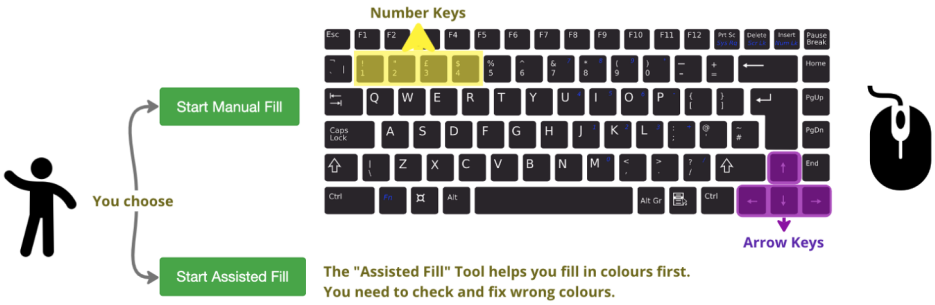

Figure 2: The experiment landing page instructing users how they can complete the paint by numbers game. There are two options for completing the task. One is Manual Fill mode, in which users fill in all colors manually through keypresses. The other is Assisted Fill, in which users wait for the simulated GenAI to fill in the colors, and then they manually fix any mistakes.

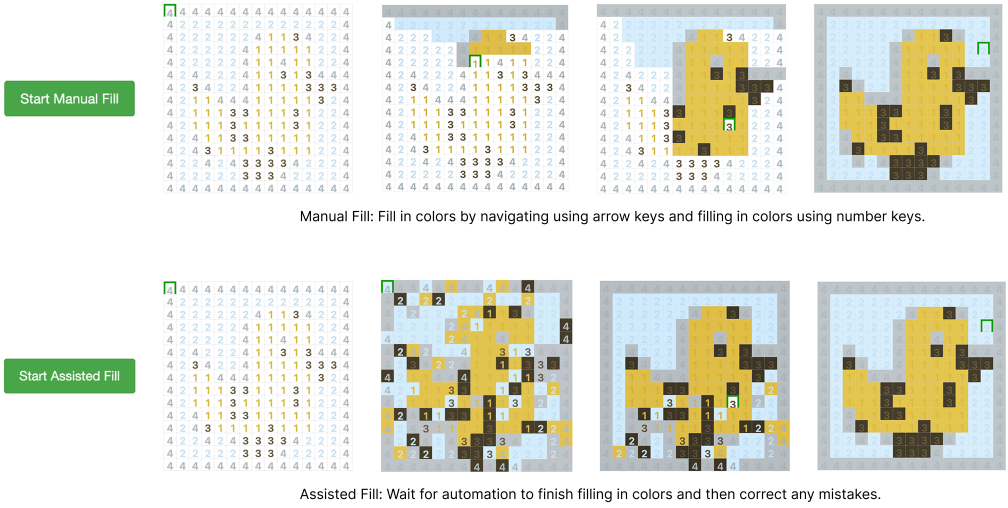

Figure 3: Illustration of the workflow for using Manual Fill (top) and using Assisted Fill (bottom). Assisted Fill simulates GenAI support tools through various latency and error rates that require users to fix any mistakes before completing the task.

random, too hard to detect, nor too apparent either. For instance, in some of the previous iterations, we tried to indicate cells that are wrong with a red border color or used the same color for both text and cell background. However, in our pilot study, we realized that these visual elements were too apparent and did not accurately mimic the subtle errors made by GenAI. We also tried using black text for all cells but found that it made identifying errors excessively challenging. The four colors used in the pixelated images were deliberately chosen to be distinct enough so that participants do not have to spend time differentiating between similar shades. The canvas size, number of cells, number of colors used and the pixilated images were all chosen through rounds of iterative design sessions and pilot studies. Our goal was for the tasks to require time and effort to complete, while also avoiding excessive cognitive or physical strain. We believe we succeeded in achieving this balance as some participants in our final study provided feedback such as "This was very very very fun!". As mentioned in Section 3.1.1, we believe this method is extensible and could also be leveraged for future studies. Specifically, this method could also be used to examine other characteristics of GenAI interaction beyond latency and error rate. Additionally, it could be adapted to compare different UI mechanisms that could help users make better decisions about when to rely on the GenAI model based on particular performance indicators, including but not limited to latency and error rate.

3.2 Experiment Design

We conducted a *between-subject* experiment to evaluate how and how well people make decisions on when to use and not use GenAI support tools. We aimed to systematically explore the full space of latency and error rate, so we assigned participants to latency conditions ranging from 0 seconds to 210 seconds with 15 seconds of step size in between (0, 15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165, 180, 195, 210) and to error rate conditions ranging from 0% to 75% with a step size of 5% (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75). For latency, we chose 210 seconds as the maximum latency based on our pilot study, which showed that 210 seconds exceeded

the amount of time participants were generally willing to wait. The step size was chosen to balance experiment size and resolution. We chose 75% as the maximum error rate because it represents the threshold for random chance. Namely, with four color options, an error rate higher than 75% would perform worse than randomly assigning colors to the cells. Similarly, the step sizes for error rate were also chosen to balance experiment size and resolution. Therefore, we ended up with 15 latency and 16 error rate conditions. We aimed to recruit 15 $latency \times 16$ $error\ rate \times 2$ $participants = 480$ participants, which would give us 15 $latency \times 2$ $participants$ + 16 $error\ rate \times 2$ $participants$ = 62 decision data points for each condition combination based on our task setup. These decision data points generated by each of our participants represented their decisions for either using or not using the Assisted Fill tool given a list of latency and error rate combinations as shown in Figure 5. Specifically, each participant was assigned one latency condition from 15 possible latency conditions. Under this fixed latency condition, they made decisions across all 16 error rate conditions, contributing 16 decision data points. Conversely, each participant was also assigned one error rate condition from 16 possible error rate conditions. Under this fixed error rate condition, they made decisions across all 15 latency conditions, contributing an additional 15 decision data points. To better explain the experiment, we present the task flow in Figure 4 with an example of one participant's performance. We will walk through the experiment and how we collected our data in detail in the next subsections.

3.2.1 Introduction and Practice. Participants were first shown an instruction through a web application, introducing the paint by numbers game, what the cells look like, the two modes of interaction and a prompt to *"try to complete the task as fast as you can!"*. Throughout the study, we kept prompting participants to complete tasks as fast as possible to maximize their productivity. We believe participants were also motivated to do so given the fixed amount of compensation, regardless of how long they took to complete the study. Participants then completed two practice tasks of filling in colors on a grid of 4 by 4 cells, one in which they used Manual Fill, and another in which they used Assisted Fill.

3.2.2 Task 1 and Task 2. After the practice round, participants completed a total of three full paint by numbers tasks with a grid size of 15 by 15. For the first two tasks, participants were assigned to complete one with Manual Fill and the other with Assisted Fill. The order of which one comes first is randomly generated. The order of the images is also randomly assigned, since in our pilot study we found no significant time or difficulty difference for tasks associated with each image. When Assisted Fill is offered, participants are also presented with the waiting time and the error rate of the tool. Participants are randomly assigned to one of the 15 latency conditions and one of the 16 error rate conditions with coverage of all combinations for the Assisted Fill tool in the first two tasks.

3.2.3 Task 3 and Survey Questions. In the final task, we informed participants that a new Assisted Fill tool is assigned to them. Whether they will be using the Assisted Fill tool or filling in colors manually depended on their answer to two questions about their tolerable waiting time and error rate for Assisted Fill given a specific error rate or wait time. The two questions were customized based on their assigned Assisted Fill condition in the first two tasks. Referring back to our example in Figure 4, since the Assisted Fill assigned to the participant in Task 2 has a latency of 150s and error rate of 35%, the first question is: *"For Assisted Fill tools, if the error rate is always 35%, but the wait time can vary between 0 second to 210 seconds: What is the longest you would be willing to wait for the Assisted Fill Tool? Remember, you are trying to complete the overall task in as little time as possible."* The second question is: *"For Assisted Fill tools, if the waiting time is always 150 seconds, but the error rate can vary between 0% to 75%: What is the highest error rate you would be willing to use for the Assisted Fill Tool? Remember, you are trying to complete the overall task in as little time as possible."*

Then based on the participant's answer to Question 1, we assign them either an Assisted Fill tool or a Manual Fill tool for Task 3. As illustrated in Figure 4, the participant was assigned an Assisted Fill tool with the same error rate, but a 60 second latency (or waiting time) for Task 3. We designed the latency of the Assisted Fill tool for the third task to be (210 seconds − latency for their previous Assisted Fill tool) to balance the completion time and avoid the case that a participant has to wait long twice in a row. If the participant's answer for Question 1 is greater or equal to 60 seconds, then they will use the Assisted Fill tool for their third task and if the participants' answer is lower than 60 seconds, they will use the Manual Fill tool. After participants completed all three tasks, we collected survey responses for the question *"How did you decide how long you are willing to wait for Assisted Fill for task 3? And how did you decide at what error rate you are still willing to use Assisted Fill for task 3? What factors influenced your decision?"* To avoid low quality responses, we required participants to input at least 50 characters into the text box. Although only participants' answers to the two questions were considered meaningful data for our study, and their performance on Task 3 was not a factor in our analysis, we still required participants to actually complete Task 3, instead of ending the experiment after they provided the answers. There were two key reasons for designing the experiment like this. First, completing Task 3 balanced the overall task completion time across participants: those who experienced longer waiting times in Tasks 1 or 2 would encounter shorter waiting times in Task 3. Secondly, having participants complete Task 3 ensured they faced the consequences of their decisions, potentially leading to more meaningful reflections and responses in the subsequent survey question.

3.3 Participants, Data and Metrics

Participants were recruited from Prolific [47]. Inclusion criteria were: no issues seeing colors, and being fluent in English. During our pilot study, completing the whole experiment took on average 18 minutes. We therefore decided to compensate each participant approximately 5.11 USD for completing the task, corresponding to 17 USD per hour. We also monitored and recorded the time participants took to complete the tasks in the final study, which aligned with our pilot study, with an average completion time of 16 minutes and a median time of 14 minutes.

We aimed for 480 participants who successfully completed the study in order to have at least two participants for each of our 15 $latency$×16 $error\ rate$ condition combinations. This would allow us to generate at least 2 $participant$ × 15 $latency$ + 2 $participant$ ×

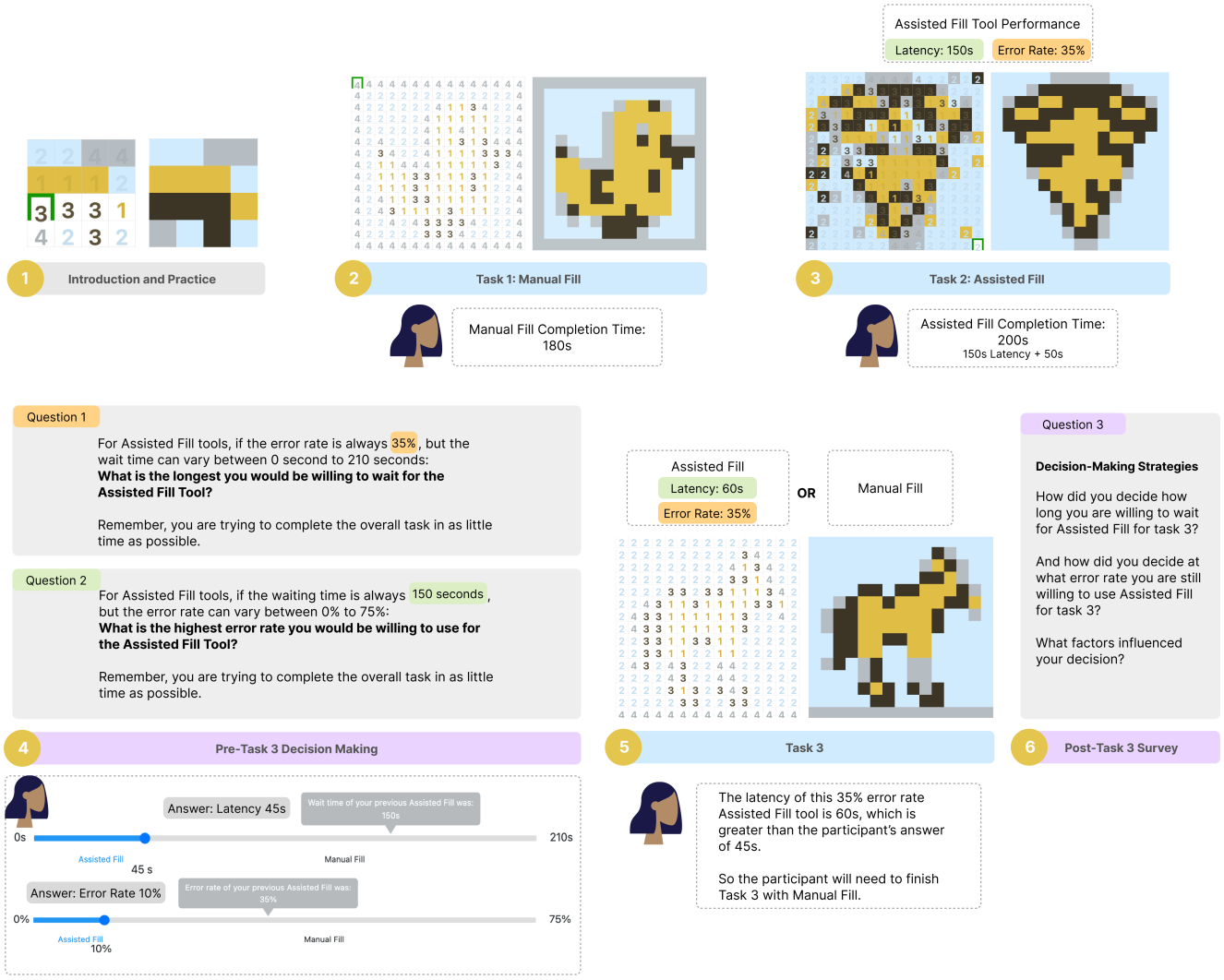

Figure 4: Workflow of the experiment and an example participant's performance. Participants went through six steps in total: (1) Introduction and two practice rounds to get to know Manual Fill and Assisted Fill; (2) The first full grid task, where participants will be assigned to either Manual Fill or Assisted Fill. In this example, the participant first was assigned to Manual Fill; (3) The second full grid task, in which participants will be assigned to a different mode of interaction from the first task; (4) Pre-Task 3 questions, in which participants answer questions that determine whether they get to use their assigned Assisted Fill tool or manually fill in colors for Task 3; (5) Task 3, in which participants complete a third paint by numbers game using the mode of interaction based on their answer to Question 1 in Pre-Task 3; (6) Post-Task 3 survey.

16 *error rate* = 62 decisions for each condition. Referring back to the participant example in Figure 4, when they chose 45s as the answer for Question 1, they were generating 1 decision data point for each of the 15 latency conditions listed (Figure 5). Similarly, when they answer question 2, they again generate 1 decision data point for each of the 16 error rate conditions. Therefore, one participant in total generates 15 + 16 = 31 data points.

We recruited 605 participants in total: 2 failed our attention test embedded in our survey questions and 95 participants were eliminated as outliers, leaving us with a final set of 508 participants. To ensure valid and consistent results, we aimed to include participants who tried their best to complete the task as fast as possible and those that could represent the average user when interacting with GenAI systems (i.e. those who were not extremely fast or slow). We detected outliers through two stages: *key gap outliers* and *total completion time outliers*. Key gap outliers are participants who have long gaps in between key presses when completing the study, indicating they may have gotten distracted. We first used the Tukey Method with k=1.5 to identify outlier (i.e. very long) key press *gaps*. Next, we used the Tukey Method with k=1.5 to identify participants who have

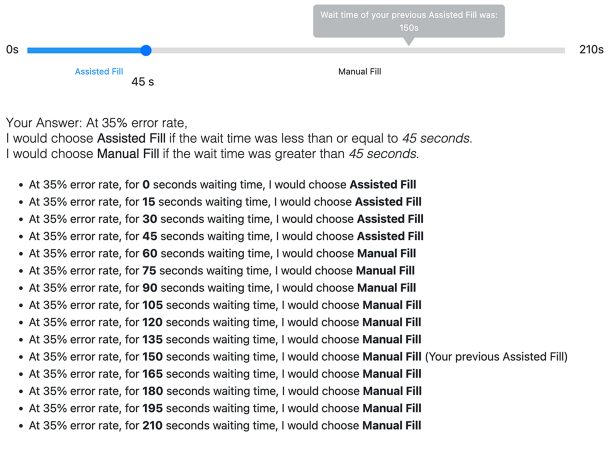

Figure 5: Participants' answers to Question 1 and Question 2. Answering one question generates 15 and 16 data points respectively.

a very high *number* of those outlier key press gaps. Since Manual Fill and Assisted Fill require different numbers of key presses, key gap outliers were analyzed by grouping all Manual Fill conditions together and grouping Assisted Fill by the different error rate conditions. Participants with an outlier number of outlier key gaps were first filtered out. Finally, to detect total completion time outliers, we again used the Tukey Method with k=1.5 and identified outliers after grouping all Manual Fill conditions together and grouping Assisted Fill conditions by the various error rate conditions. After removing outliers, we ended up with 508 participants, resulting in a total of 508×15 *latency* $+ 508 \times 16$ *error rate* $= 15,748$ decision data points, and at least 62 data points for each combination of latency and error rate.

3.3.1 Optimal Answer and Break-Even Value for Latency and Error Rate. Answers for Question 1 and 2 provided us with participants' actual choices of when to use or not to use an Assisted Fill tool. This data is referred to as a participant's actual choice or their chosen answer. We also generated participants' *optimal* answer (or *optimal* choice) through their completion times for Task 1 and Task 2 using Manual Fill and Assisted Fill. Participants' optimal answer is determined by the break-even value, calculated based on their performance in the previous two tasks. Specifically, if the latency or error rate of the Assisted Fill tool is below the break-even value, participants will complete the task faster using the tool. However, if the latency or error rate exceeds the break-even point, participants would be better off completing the task manually.

Referring back to the example in Figure 4 (step 3), this participant used 50s to fix the mistakes generated by the Assisted Fill tool with a 35% error rate. Therefore, the participants' *optimal* answer for Question 1 around their longest tolerable latency would be:

180s (Manual Completion Time) – 50s (Time To Fix Errors) = 130s

So, after waiting 130s for Assisted Fill to finish, the user can still complete the task faster than manually filling in all colors. In our example, the participant did not choose optimally, because their chosen latency of 45s for an Assisted Fill tool with 35% accuracy is much shorter than their optimal latency of 130s. As for Question 2, given an Assisted Fill tool with 150s latency, then to ensure that the participant completes the task with the Assisted Fill tool faster than filling it in manually, they would only have 180s (Manual Completion Time) - 150s (Given Latency) = 30s available for fixing any mistakes. On average, this example participants' speed (in cells per second) for fixing the Assisted Fill tool's mistakes was:

$$\frac{35\% \text{ Error Rate} \times 225 \text{ Total Number of Cells}}{200\text{s Assisted Completion Time} - 150\text{s Latency}} = 1.58 \text{ cells / s}$$

Therefore, with 30 seconds given to fix mistakes, the participant could fix $30\text{s} \times 1.58$ cells / s $= 47.4$ cells, which corresponds to an optimal value of 47.4 / 225 = 21% for the error rate. Any error rate larger than 21% would cause the participant's Assisted Fill completion time to be longer than their manual completion time. As with their choice for latency, the participant in Figure 4 did not choose optimally. They chose a highest tolerable error rate of 10%, which is below their optimal answer for error rate. At 10% error rate (< 21%), they would be faster using Assisted Fill. Based on this method, we were able to calculate optimal answers for latency at a given error rate and error rate at a given latency for each participant and compare them with their actual choices.

3.4 Pilot Study

We conducted a series of pilots involving the authors of the paper as well as five other participants from our personal network. The pilot sessions were helpful in designing the system for our study. As discussed in Section 3.1, through the pilot sessions, we revised our UI design and experiment design to make sure the system effectively emulated key characteristics of user interactions with GenAI and to streamline the experiment to provide a smooth and intuitive experience for participants. Beyond the color and cell number of the task design, some other UI elements that were redesigned through the pilot studies include a warning message and alert sound pop up when users leave or click outside of the experiment web page. The

slider in combination with a list of answers corresponding to the selected value in the slider was also a design consideration coming out of the pilot sessions to enhance participants' understanding of their choice (as seen at the bottom of Figure 5).

4 Results

4.1 RQ1: Quantitative Analysis of Actual Choice vs. Optimal Choice

4.1.1 Given an error rate, what latency is tolerable? For each of the 16 error rate conditions (0%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%), we were able to plot out two curves based on participants' answers to Question 1: Given a particular error rate, what is the longest time they are willing to wait, and what is their optimal waiting time. By aggregating all participants' choices for each latency, we obtain the percentage of participants choosing to use Assisted Fill at a given latency and error rate as well as the percentage of participants who *should* be using Assisted Fill if they were using an optimal strategy (Section 3.3.1). Sigmoid curves were fitted to the data, plotted along with 1-sigma bootstrap confidence intervals [8, 40].

Among the set of 16 plots, we found two patterns which could be represented by the top left and the bottom right line graph in Figure 6. They represent the percentage of participants choosing Assisted Fill with $y = 1$ meaning everyone chose Assisted Fill and $y = 0$ meaning no one chose Assisted Fill. The x-axis represents the different latency conditions, while the four charts show increasing error rates: 0%, 20%, 50%, and 75% respectively. Transition stages are represented through two plots on top right and on bottom left, illustrating how the figure representing error rate at 0% eventually becomes the figure at 75%. Namely, as error rate increases, the area of the gulf of impatience becomes smaller and the gulf of overreliance becomes larger. As illustrated in Figure 6, given an error rate of 0%, participants tend to underestimate Assisted Fill's support when latency is low, creating a gulf of impatience. On the other hand, when latency gets high, users tend to overestimate Assisted Fill's support, creating a gulf of overreliance. As the error rate increases, we see through a series of transition stages that the gulf of impatience gets smaller and gulf of overreliance gets bigger.

By calculating the differences between each of the 16 pairs of line plots and aggregating the differences between the actual choices and the optimal choices, we are able to obtain the heat map shown in Figure 7. This illustrates that people generally make good decisions when Assisted Fill is either perfect or extremely bad. However, participants tend to underuse Assisted Fill when the performance is moderate in both error rate (5% - 15%) and latency (25s - 100s). This is indicated by the red area on the bottom left of the figure, labeled as the gulf of impatience. Conversely, as shown by the blue area in the bottom right part of the heat map, participants over-rely on the Assisted Fill tool when it has a high error rate (50% - 75%) and low latency (0s - 25s), which results in a gulf of overreliance.

4.1.2 Given a latency, what error rate is tolerable? When participants were given a specific latency and asked for what is a tolerable error rate, we found similar trends. For each of the 15 latency conditions (0s, 15s, 30s, 45s, 60s, 75s, 90s, 105s, 120s, 135s, 150s, 165s, 180s, 195s, 210s), we were again able to plot out two discrete curves based on participants' answers to question 2. Given a particular latency, a curve that represents the largest error rate participants are willing to tolerate for Assisted Fill, and a curve based on their optimal choices. Through aggregating participants' choices, we are able to plot one line illustrating the percentage of participants choosing to use Assisted Fill at each latency and error rate condition as well as the percentage of participants who should be using Assisted Fill if they chose optimally. Again, a sigmoid curve was fitted for each line plot to provide a clearer representation, along with 1-sigma bootstrap confidence intervals [8, 40].

As shown in Figure 8, we found that given low latency (<30 seconds), participants tend to underestimate their productivity when supported by automation with almost all error rate conditions, resulting in a large gulf of impatience. When waiting time is between 30 seconds to 60 seconds, participants tend to underestimate Assisted Fill with high performance (error rate between 0% and 40%) and over-rely on low performance Assisted Fill tools (error rate greater than 40%). As the given latency increases, the gulf of impatience becomes smaller and the gulf of overreliance grows.

Through calculating the differences between each of the 15 pairs of line plots (4 of which are represented in Figure 8) and aggregating the differences, we again obtain a heat map shown in Figure 9. Similarly, this figure illustrates that people generally make good decisions when Assisted Fill is either perfect or extremely bad. However, participants tend to underuse Assisted Fill when the performance is moderate in both latency (0s - 30s) and error rate (10% - 50%), as shown through the gulf of impatience label in the red area. Conversely, as shown by the blue area in the right part of the heat map, participants over-rely on Assisted Fill tools when it has a high latency (120s - 210s) and low error rate (0% - 10%), resulting in a gulf of overreliance.

4.1.3 Combining and Locating the Gulfs. Combining answers for Question 1 and Question 2 together provide us with at least 62 data points (15 X 2 + 16 X 2 = 62) for each of the 15 latency and 16 error rate conditions. Visualizing the gap between participants' actual choices and their optimal choices through the heat map in Figure 10 presents us with a clearer view of where the gulf of impatience and the gulf of overreliance are located. To summarize, we found that people make optimal decisions when Assisted Fill is either high performing in both latency and error rate or low performing in both latency and error rate. In other words, people make good decisions on deciding whether to use automation or not when encountering "smart AI" and "dumb AI." The gulf of impatience occurs when the AI's performance becomes more moderate. The gulf of overreliance occurs especially when one of the two conditions is extremely high performing and the other is low performing, which is indicated through the dark blue color in the top left corner as "smart but slow AI" and the bottom right corner of the graph as "dumb but fast AI".

4.2 RQ2: Qualitative Analysis of Strategies Used in Decision Making

Participants articulated various strategies they used to make their decisions for when to use and not to use Assisted Fill when trying to complete the task as fast as possible. One author went through all responses to generate initial codes representing strategies that participants used to support their decision-making. Then all authors

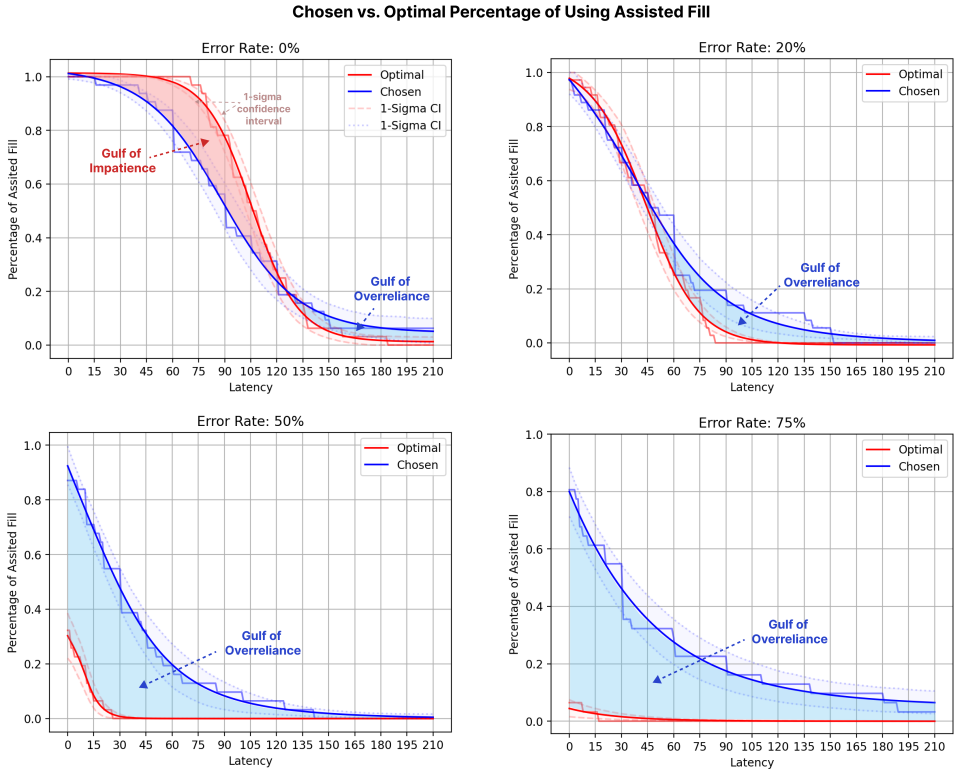

Figure 6: Four plots illustrating the main patterns of chosen versus optimal percentages of using Assisted Fill given various error rate conditions. The y-axis in each chart represents the percentage of participants choosing Assisted Fill, with $y = 1$ meaning everyone chose Assisted Fill and $y = 0$ indicating no one chose Assisted Fill. The x-axis represents the different latency conditions. The charts represent these variables for increasing error rates: 0% (top left), 20% (top right), 50% (bottom left), and 75% (bottom right). Sigmoid curves are fitted to the raw data to illustrate trends. Shaded areas around the red and blue curves represent 1-sigma bootstrap confidence intervals [8, 40].

came together for several rounds of discussion to iteratively discuss the codes and come to an agreement. Through further discussions, we then synthesized these into three key strategies: (1) calculating time; (2) relying on "gut feeling" thresholds; and (3) balancing fun and effort. Some participants explicitly mentioned one strategy, while many others used multiple strategies. As shown in Figure 11, which of the three strategies participants did or did not use had little influence on how effectively they made their decisions.

4.2.1 Time calculation. Many participants made their decisions based on their estimation of how long it took them to complete the task using the two different modes of interaction. We found that 339 participants (66.7%) explicitly mentioned that their answers for Question 1 and Question 2 were calculated based on estimating how fast they complete the first two tasks and how fast they could fill in colors and fix mistakes. To illustrate how participants described their strategy, we showcase three examples:

- *"To decide how long to wait, I tried to remember how long it took me to complete the task manually. I tried to compare the time it took me to do the whole task from scratch (manually) to when the waiting period for the Assisted Fill option would end up hurting my time rather than helping me. For the Assisted Fill, around half of them are filled correctly, so if I do not wait for too long, then I can beat the time I took to do the Manual Fill since I have less to fix. For the error rate, I again tried to make it so that I could finish the task as soon as possible. I knew that the [waiting] time of 195 seconds was a very long time (it felt that way to me at least), and I thought I could finish the task manually within the 195 seconds so I wanted to choose the lowest error rate since it meant less to fix."*
- *"I checked in how much time I did the first task manually and compared it to the time I'm willing to wait for autofill [Assisted Fill] to work. Because the aim was to complete the task as fast as possible, I had to select a waiting time that is less than the time I would have taken to do the task manually and leave enough time to correct mistakes as well."*
- *"I tried to think how much time I would take doing it manually and what time it would compensate if I use Assisted Fill. [For] error rate, I needed to be very low because otherwise I would have to do almost everything manually after."*

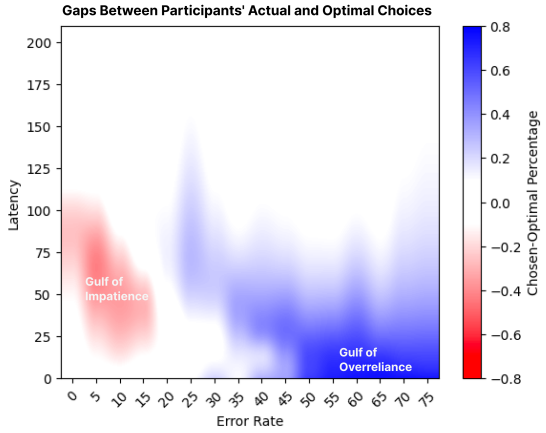

Figure 7: A heat map generated with 16 pairs of sigmoid curves (Figure 6) across 16 error rate conditions, illustrating the differences between the percentage of participants' actual choices of choosing Assisted Fill and the percentage of participants' optimal choices of choosing Assisted Fill. White areas indicate close to optimal choices, blue areas show overreliance on Assisted Fill, and red areas show underreliance on Assisted Fill. Data used here was collected from participants' answer to Question 1 before working on Task 3: Given X error rate, what is the longest time a participant is willing to wait for the Assisted Fill tool.

4.2.2 "Gut feeling" threshold. The second popular strategy is making decisions based on a "gut feeling" derived from previous experiences, which appeared in 39.4% percent of the quotes. Participants who used this strategy did not mention that they estimated time or speed, but emphasized that any waiting time or error rate above a certain threshold would not be tolerable. For instance, participants mentioned *"I decided I am willing to ask for 45 seconds because, I simply don't want to wait more"* and *"I thought that 30 seconds is the maximum time that a person won't be bored and not focused."*

Some participants derived their threshold from other experiences beyond the study. For example, one participant referred to their experience watching YouTube advertisements, *"even on YouTube the max ad length is 30 seconds long (I guess, I use Premium) but it is yet tolerable. And I don't want to wait a long time for something that is not even entirely useful to me."* Other participants referred to "regular waiting time" and "acceptable error rate" as the basis for their decision-making. One mentioned *"I think we live in a society where time is important, people don't want to be idle for too long, and due to the internet, people's attention span is relatively short. Because of that, I'd rather not wait too long without doing anything, and if I had to wait longer for Assisted Fill to help me with a task, I would prefer it to be worthwhile. Basically I'd want the error margin to be worth the time I waited, otherwise I could just do it myself. Anything more than a 30% chance of error would not be worth my time."*

For error rate, some participants referred to their expected performance of state-of-the-art AI and mentioned that they would not use Assisted Fill unless it is on par with their expectation. For example, one participant mentioned: *"I thought that nowadays tools like these can take a maximum of a pair of seconds to have this kind of filling realized"* and similarly, another participant mentioned: *"Given the range of error[s] that was presented to me, 75% was way too much of a risk to take to have to wait around for an AI (I know it's not an AI, it's just for lack of a better term) to complete the task. I would not be willing to wait more than 5 seconds for it to complete what are very simple inputs."*

4.2.3 Fun and effort. Even when reducing time of completion was stressed to be the main goal of the task, 26.6% of our participants mentioned fun and effort as factors affecting their decision making processes. While some participants found that filling in colors manually is more fun, others found fixing colors from Assisted Fill more fun. Some participants who enjoyed Manual Fill mentioned:

- *"I actually found more enjoyment from doing it manually."*
- *"I'm impatient and I actually enjoyed it, and found this paint by numbers therapeutic…And frankly it frustrated me to have to review what the machine did in this specific case."*
- *"In Manual mode, after a couple rows my brain and muscle memory started to really help make the task easier, and thus faster. And last but not least: Manual mode the filling in felt like a game, a rhythm game to be exact, where it's really satisfying to get as many right as you can without missing "a beat" in the string! :) (And I'm not generally someone who likes rhythm games at all.) Automatic mode really did not feel satisfying."*

On the other hand, many other participants mentioned that Assisted Fill is more fun. For example:

- *"Correcting the auto-fill [Assisted Fill] mistakes wasn't bad at all so I didn't mind a higher mistake rate."*
- *"I think it's much easier to not notice the wrong colors when you use the associated tool, when you do it manually it's almost impossible to get confused about it."*
- *"I found Assisted Fill more fun so I would choose this, it was more pleasant (error rate 60)"*
- *"Half full is easier than starting from scratch."*

Not surprisingly, waiting was mentioned to be "boring" for a few participants, which influenced their decision making. For example,

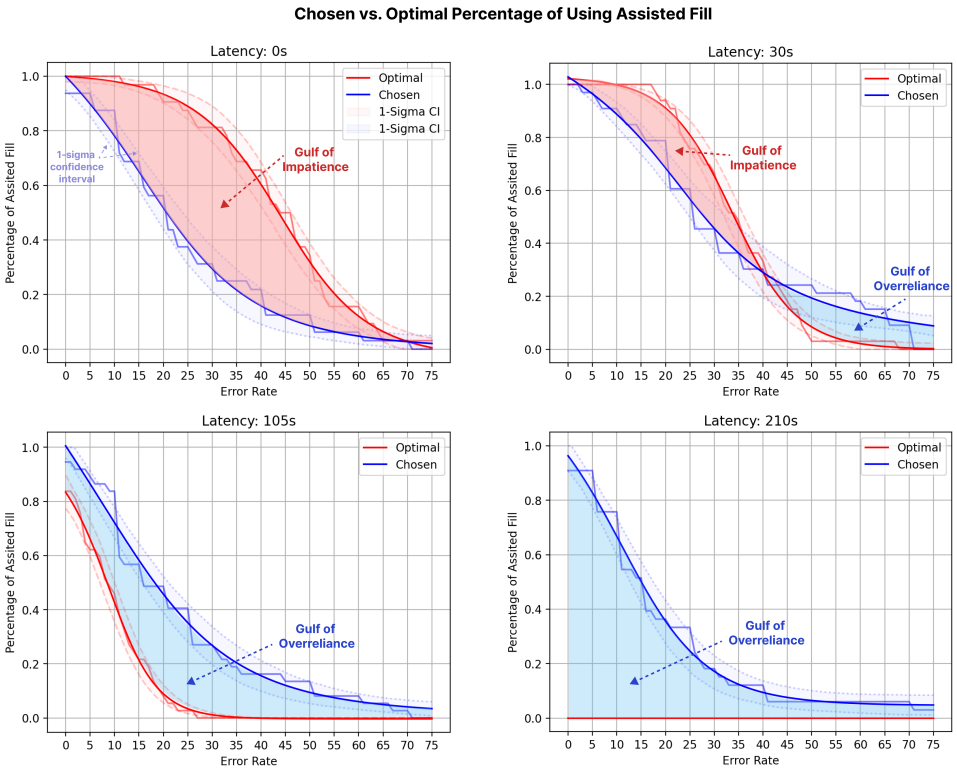

Figure 8: Four plots illustrating key patterns in chosen versus optimal percentages of using Assisted Fill given various latency conditions. The y-axis in each chart represents the percentage of participants choosing Assisted Fill, with $y = 1$ meaning everyone chose Assisted Fill and $y = 0$ indicating no one chose Assisted Fill. The x-axis represents the different error rate conditions. The four horizontal charts represent these variables for increasing latency: 0s (top left), 30s (top right), 105s (bottom left), and 210s (bottom right). Sigmoid curves are fitted to the raw data to illustrate trends. Shaded areas around the red and blue curves represent 1-sigma bootstrap confidence intervals [8, 40].

one mentioned: *"I went based on the time I felt I could tolerate waiting before getting bored"*. Another participant shared: *"I'm a very impatient person, so I prefer to do things myself more than waiting."*

5 Discussion

5.1 Decision-making Support Beyond Presenting AI's Performance

Understanding humans' decision making process of when to use or not use an automation is not a new topic in HCI. As early as 1983, Bainbridge [1] pointed out that designers of automated systems often leave human operators with a collection of arbitrary tasks and offer little thought to providing support for them, resulting in failed human-computer collaboration and "ironies of automation."

In recent years, human-AI interaction researchers have also recognized that merely developing better models does not lead to effective collaboration between humans and automated systems [39, 53]. To better facilitate human-AI interaction, researchers highlighted the importance of providing information on model performance and explanations to facilitate appropriate reliance on and trust in AI systems [64, 66, 68]. Our study illustrates that having information about GenAI support tools' performance does not guarantee users will make better decisions on when to use or not use GenAI. This finding also echoes previous HCI research by Kloft et al. [24], illustrating that people are biased in their assessment of an AI's performance even when they were informed by verbal descriptions of the AI. Specifically, they found that people judged performance with AI-assistance as superior, even when in fact the AI used in the experiment was nonfunctional. In algorithmic decision-making in healthcare, Cao et al. [4] found that people's reliance on an AI's answer is impacted by how model uncertainty is communicated.

Extending prior studies on binary decision-making with AI, our study identifies and characterizes the gulfs of underreliance and overreliance that occur when people are presented with different waiting times and error rates in the context of human-GenAI collaboration. Interaction with GenAI has unique characteristics and challenges, as it often involves open-ended solutions, unpredictable outputs, and iterative workflows where generated results serve as intermediate outcomes that users must refine and correct to achieve their ideal results. **Our findings highlight that merely**

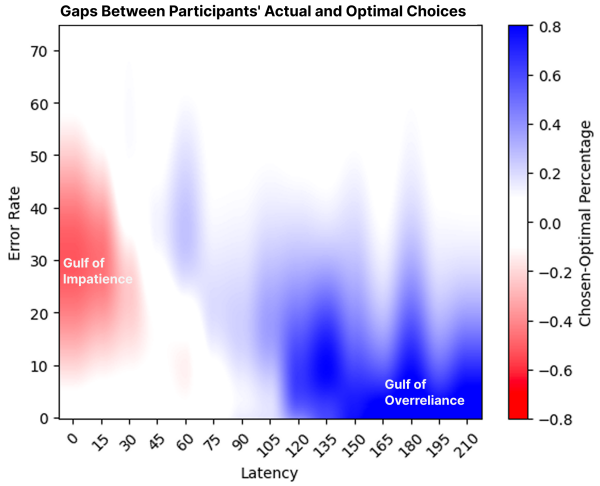

Figure 9: A heat map generated with 15 pairs of sigmoid curves (Figure 8) across 15 latency conditions, illustrating the differences between the percentage of participants' actual choices of choosing Assisted Fill and the percentage of participants' optimal choices of choosing Assisted Fill. White areas indicate close to optimal choices, blue areas show overreliance on Assisted Fill, and red areas show underreliance on Assisted Fill. Data used here was collected from participants' answer to Question 2 before working on Task 3: Given X latency, what is the largest error rate a participant is willing to tolerate for Assisted Fill.

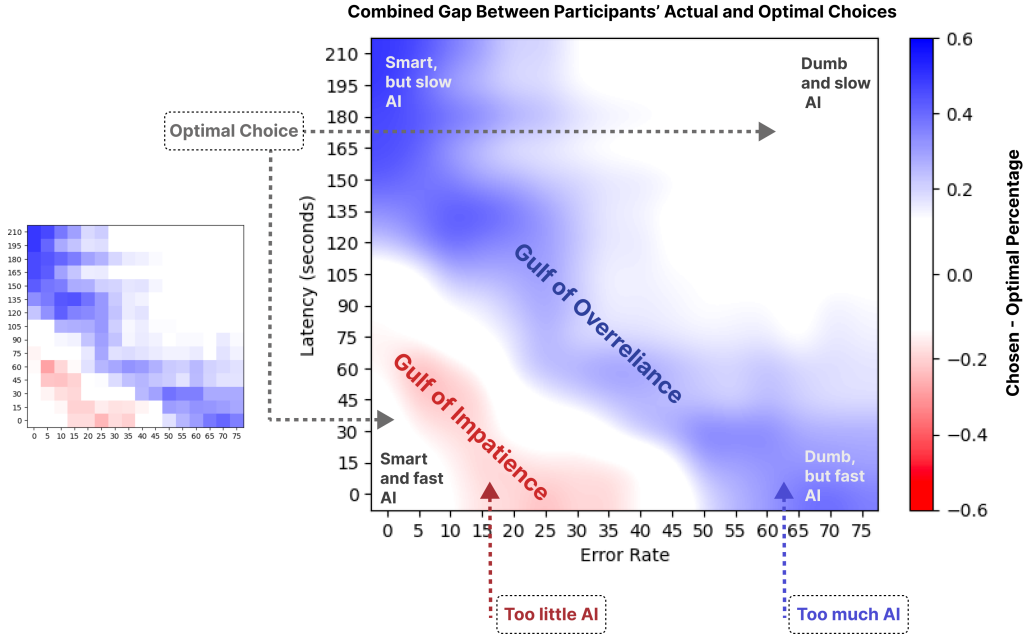

Figure 10: A heat map based on the gap between participants' answers for Question 1 and Question 2 and their optimal choices. We obtained this continuous heat map by applying bicubic smoothing to the discrete heat map on the left. White areas indicate close to optimal choices, blue areas show overreliance on Assisted Fill, and red areas show underreliance on Assisted Fill.

presenting information about the model's performance is not enough in helping users make optimal decisions about whether or not to use a GenAI productivity support tool. To help bridge the gulfs of impatience and overreliance in human-GenAI interaction that we identified, we offer the following three design implications that designers can consider:

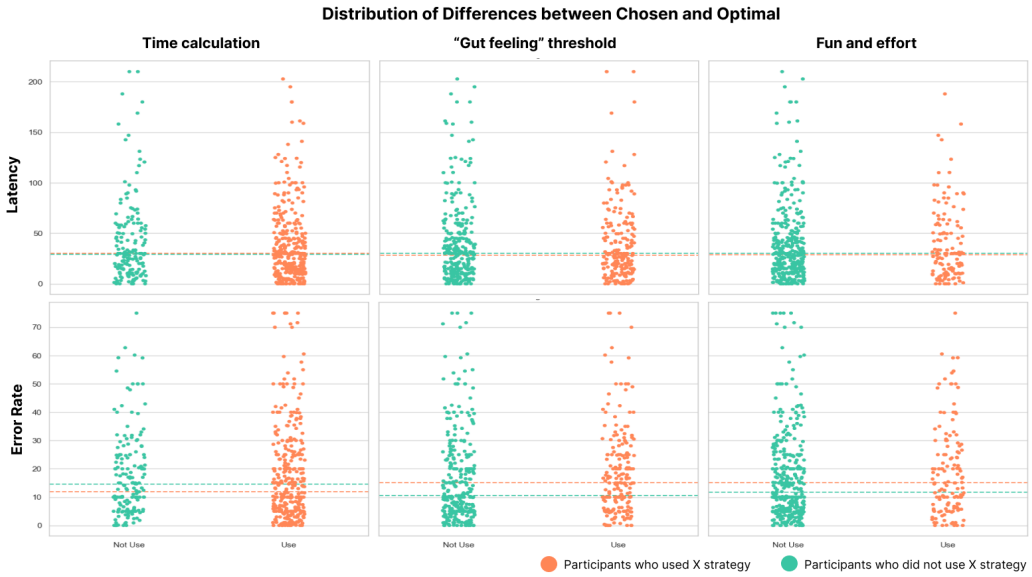

Figure 11: Strip plots illustrating the distribution of differences between chosen and optimal latency (top) and error rate (bottom) by adoption of strategy. The orange strip plots illustrate the distribution of participants who leveraged a certain strategy and the green strip plot illustrate the distribution of participants who did not leverage that strategy. The dashed lines represent the medians of each group. There is not much difference across each pair.

- **DI1:** Collect user performance data, so systems can calculate and respond to individual behavior patterns.
- **DI2:** Leverage UI design techniques to calibrate impatience and overreliance.
- **DI3:** Design for non-blocking AI in suitable scenarios to avoid suboptimal decisions.

5.1.1 DI1: Collect User Performance Data. For our study, we designed a system that collects participants' performance data, enabling us to calculate participants' optimal choices of when to use or not use the Assisted Fill tool. This illustrates the value of collecting user performance data when designing GenAI systems for productivity support. For supporting tasks containing processes that are repetitive in nature, like 3D modeling, systems could gather performance data, estimate the impact of the GenAI tools, and respond to users' behavioral patterns to address suboptimal decisions. For instance, Mozannar et al. [39] have developed a system for programmers to retrospectively document how much time they spend on each type of common activity when interacting with Copilot. Future research could explore automatic documentation of users' performance on various tasks with and without GenAI's support.

5.1.2 DI2: Leverage UI Design Techniques. In line with Norman's paper on appropriately designing for automation [41], recent research [5, 53] stressed the importance for GenAI tools to continuously provide relevant feedback to users about the system's state. We suggest that researchers and designers attend to interface design, learning from researchers studying earlier generations of computing systems including progress bars, loading screens, and countdown interfaces. These elements can be leveraged to calibrate users' perceived waiting time, trust and reliance [10, 15, 18, 25, 55]. For instance, Harrison et al. [15] found that progress bar with ripples heading left makes progress appear faster than it really is, and more recently Komatsu et al. [25] found that people's perception of time can be impacted by the interface design of countdown and countup screens. Therefore, when systems collect users' performance data and know users' tendency to be either impatient for AI or overreliant on AI, such UI design techniques could be strategically leveraged to nudge users towards the optimal choice in that situation and help bridge decision-making gaps.

5.1.3 DI3: Design for Non-Blocking GenAI. While not universally applicable, there are certain scenarios in which it is possible to completely avoid having users decide between using or not using GenAI. This is another strategy that designers can employ to reduce users' cognitive burden and prevent suboptimal decisions. Concretely, we recommend that system designers seek ways to embed GenAI in non-blocking tasks or have GenAI operate in the background. Non-blocking GenAI tasks reduce the the weight of the decision of whether or not to use GenAI and its impact on the user's productivity. While the GenAI operates asynchronously in the background, the user can engage with other subtasks while waiting for the GenAI system to complete. Since users can still make progress while waiting for the AI to complete, impatience or underreliance may be less of a concern as the user can keep themselves busy with other useful tasks, and overreliance on the AI may also be less problematic since this suboptimal choice may have much less impact on the users' overall productivity. For certain tasks, it may even be possible to distribute the work between the GenAI system and the user. The user could be handed tricky edge

cases that the GenAI has difficulties handling, while the GenAI works away on less complex aspects of the task. In these situations, the user and GenAI could work in parallel on the same task in a mixed-initiative way [19], with smart merging of the results from the GenAI and from the user. In summary, we recommend designers to seek ways of embedding GenAI in the background whenever feasible, particularly where tasks can be non-blocking. This approach minimizes the number of factors users must consider, simplifying their workflow, and mitigating risk of suboptimal decision-making.

5.2 Task Delegation for Different Types of Imperfect AI

Desired roles of AI in human-AI collaboration have been discussed from different perspectives such as roles in creativity processes [36, 69], ideation processes [37] and data storytelling [29]. Zhou et al. [69] also argued diverse AI roles may need to be implemented simultaneously to fulfill users' diverse needs.

Categorizing AI through the focus on reliance and productivity, Figure 10 illustrates four distinct types of AI agents – "smart AI", "dumb AI", "smart but slow AI", and "fast but dumb AI". Beyond illustrating how well people make decisions around when to use or not use GenAI given different latency and error rate conditions, Figure 10 also sheds light on people's preferences on collaborating with different types of AI agents. We found that users are great at embracing the "smart AI" and avoiding the "dumb AI." Even though researchers hope to advance AI so that all AI tools eventually have minimal latency and error rate, given limited resources, understanding how people work with less than perfect AI is at the moment more valuable in supporting collaboration, productivity, and satisfaction in human-AI interaction.

As shown in Figure 10, people tend to over-rely on "smart but slow AI" as well as "fast but dumb AI," which echoes past research in trust [28] and cognitive offloading [49], and suggests potential design opportunities for affording task delegation to different types of AI assistance. High accuracy AI that may be slow to produce output (i.e. "smart but slow AI") still provides users a sense of trust for completing the task that requires precision with high quality results. Past research found that trust is especially important for navigating uncertainty [28], which is a key challenge of interacting with GenAI. Therefore, as people prioritize trust to avoid uncertainty and complexity, they might prefer to work with "smart but slow AI" for tasks that require precision, even if it might be at the cost of their productivity, so that there is more time available to focus on more strategic and creative tasks. Overreliance on "fast but dumb AI" echoes research in cognitive psychology that suggests that people seek ways to offload mental work. A "fast but dumb AI" can provide a quick rough draft solution, allowing people to bypass the initial, often most mentally taxing, part of a task quickly [49]. Thus people might have the tendency to also over-rely on "fast but dumb AI" to start the initial work and then correct their mistakes, even if the total completion time might be longer than doing everything on their own. Therefore, we believe that a solid understanding of different types of imperfect AI is valuable for advancing human-AI collaboration given limited resources. Previous work on the desired quality of fabrication automation found that people hope to be able to negotiate trade-offs between time, quality, and cost [65]. Our study aligns with these findings and shows the potential of enabling users to delegate tasks to different types of AI agents, or perhaps even allowing systems to delegate tasks on the user's behalf in certain scenarios.

To effectively leverage task delegation in human-GenAI workflows, we see a couple of promising opportunities for future research. One avenue is exploring how users rely on different types of imperfect AI for different tasks and scenarios. We propose that future studies investigate how users interact with GenAI systems across multiple tasks within a single workflow. These studies could focus on how users' reliance on AI would shift when interacting with different types of AI imperfections. For instance, we hypothesize that users would be less sensitive to waiting time and more inclined to prefer "slow but smart AI" for non-blocking tasks, as opposed to blocking tasks that require immediate feedback. Secondly, we also see potential in better understanding how delegation of different tasks to different types of imperfect AI would impact users' performance and satisfaction. In our study, we could already start to infer some preferences for different types of imperfect GenAI (Figure 10). Recent work has started to understand users' performance and preferences for AI-initiated delegation in decision-making tasks [16, 17, 23]. Future research could look into delegating various tasks that are involved in human-GenAI interaction.

5.3 Limitations and Future Work

We recognize that our study comes with some limitations. We explored one type of task and one type of GenAI assisted tool using a simulated environment, and prioritized controllability (or precision) in our online experiment. GenAI tools in real-world applications are different from a simulation, and this combined with the controlled nature of the study may have come at the expense of realism [38, p. 155]. However, as discussed in our methodology section, conducting a controlled experiment like ours presents "both opportunities for gaining knowledge and limitations to that knowledge" [38, p. 154]. Additionally, not all tasks supported by GenAI tools require high accuracy like coding, 3D modeling, or the task that we simulated in our experiment. For other tasks, GenAI support tools such as text-to-image models or Large Language Models could be extremely useful for brainstorming during early stages of design or communication processes of design [20]. In those scenarios, output manipulation might not be required and productivity and efficiency could be understood differently. Shi and Deng [52] also showed that users might have different preferences on when and how latency or delay is communicated when interacting with AI-enabled tools.

Therefore, we hope to inspire further research to explore how people make their decisions around when to use or not to use GenAI for different types of scenarios and modes of interaction, which we will discuss in the next subsections.

5.3.1 Extending the Methodology to Other Types of Tasks and Tools. As discussed above, people are leveraging GenAI for a diverse range of tasks. Our experiment focused on people's decision-making for blocking tasks, where users have to wait for the GenAI tool to finish before moving on to the next steps. However, GenAI is also being used for non-blocking tasks, where users can continue working on other tasks while waiting for GenAI to complete an intermediary output. Future HCI research could look into decision-making for

GenAI-supported tasks that are non-blocking and ways to support users in these scenarios. Another interesting opportunity would be to extend our method to study non-blocking GenAI. Researchers could also leverage our methodology to vary other conditions like difficulty in identifying errors and UI design for nudging people to make more optimal decisions, and extending to other automation-assisted decision-making tasks beyond GenAI.

5.3.2 Long-Term Versus Short-Term Use. We studied how people make choices within a short period of time. However, people learn from past experiences and could eventually get better at making decisions on whether to use or not to use GenAI under different circumstances. Therefore, we see value in conducting a longer experiment or longitudinal study on how people's decision processes change through their interaction with generative AI support tools. This avenue of exploration was also explicitly pointed out by one of our participants by stating "I hoped to try them for longer."

5.3.3 Judgment and Decision Making. Past theories in judgment and decision-making [22] also become critical when bringing our experiment results to real-world situations. Scholars in HCI and behavioral psychology [21, 60] have long emphasized the importance of recognizing sociotechnical factors beyond rationality. As GenAI support tools are no longer limited to supporting completion of a specific type of task, but shifting how human work takes place [56, 67] and impacting how teams collaborate [14, 56], it is more and more important to acknowledge the complexity of judgment and decision-making. As early as the 1980s, Lucy Suchman's influential work on AI research at the time, called for a shift from "plan" to "situated action," recognizing that people's decision-making processes go beyond rationalistic calculation and execution [54]. This is echoed by our findings. Participants described relying on various strategies to make their decisions. Thus, we believe longitudinal studies and ethnographic research beyond controlled experiments could be very valuable for human-AI interaction.

6 Conclusion

In this paper, we investigated people's decision-making and reliance on GenAI support tools for productivity enhancement. Our controlled experiment illustrated the potential of creating a controlled environment for systematically testing specific variables representing characteristics of GenAI. Our study varied the latency and error rate of a simulated GenAI support tool, and revealed key patterns in how users make or fail to make optimal decisions when deciding whether to rely on GenAI automation. We identified and located the "gulf of impatience" and the "gulf of overreliance", extending the concepts of overreliance and underreliance in ML-assisted decision making to the context of using GenAI to maximize productivity. We highlighted three key strategies participants adopted when making their decisions. While no particular strategy led to more optimal decision making, our study illustrated the entanglement of factors affecting participants' decision-making processes. From our findings, we distilled three design implications to better support users in making informed decisions about whether to use GenAI productivity tools, and suggest opportunities for leveraging task delegation when interacting with various types of imperfect GenAI.

References

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

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

[3] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258–262. https://doi.org/10.1145/3301275.3302289

[4] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–32. https://doi.org/10.1145/3637318

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

[6] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[7] Stephen R. Dixon, Christopher D. Wickens, and Jason S. McCarley. 2007. On the independence of compliance and reliance: are automation false alarms worse than misses? *Human Factors* 49, 4 (Aug. 2007), 564–572. https://doi.org/10.1518/001872007X215656

[8] Pierre Dragicevic. 2015. *HCI Statistics without p-values*. Research Report RR-8738. Inria. 32 pages. https://inria.hal.science/hal-01162238

[9] Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3613904.3642782

[10] Gustavo Machado de Freitas, Natan Luiz Paetzhold Berwaldt, Gabriel Di Domenico, Alfredo Cossetin Neto, and Cesar Tadeu Pozzer. 2024. Interactive 2D vs. 3D Loading Screens in VR: Impact on User Experience and Perceived Wait Time. In *Proceedings of the 26th Symposium on Virtual and Augmented Reality (SVR '24)*. Association for Computing Machinery, New York, NY, USA, 122–127. https://doi.org/10.1145/3691573.3691585

[11] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–29. https://doi.org/10.1145/3613904.3642002

[12] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

[13] Frederic Gmeiner, Humphrey Yang, Lining Yao, Kenneth Holstein, and Nikolas Martelaro. 2023. Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3580999

[14] Yuanning Han, Ziyi Qiu, Jiale Cheng, and RAY LC. 2024. When Teams Embrace AI: Human Collaboration Strategies in Generative Prompting in a Creative Design Task. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3613904.3642133

[15] Chris Harrison, Zhiquan Yeo, and Scott E. Hudson. 2010. Faster progress bars: manipulating perceived duration with visual augmentations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1545–1548. https://doi.org/10.1145/1753326.1753556

[16] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. 2022. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts, Vol. 3. 2478–2484. https://doi.org/10.24963/ijcai.2022/344 ISSN: 1045-0823.

[17] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 453–463. https://doi.org/10.

1145/3581641.3584052

[18] Jess Hohenstein, Hani Khan, Kramer Canfield, Samuel Tung, and Rocio Perez Cano. 2016. Shorter Wait Times: The Effects of Various Loading Screens on Perceived Performance. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3084–3090. https://doi.org/10.1145/2851581.2892308

[19] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[20] Rong Huang, Haichuan Lin, Chuanzhang Chen, Kang Zhang, and Wei Zeng. 2024. PlantoGraphy: Incorporating Iterative Design Process into Generative Artificial Intelligence for Landscape Rendering. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–19. https://doi.org/10.1145/3613904.3642824

[21] Edwin Hutchins. 1995. How a cockpit remembers its speeds. *Cognitive Science* 19, 3 (July 1995), 265–288. https://doi.org/10.1016/0364-0213(95)90020-9

[22] Gideon Keren and George Wu (Eds.). 2016. *The Wiley Blackwell Handbook of Judgment and Decision Making, 2 Volume Set* (1st edition ed.). Wiley-Blackwell, Chichester, West Sussex.

[23] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2021. Towards Unbiased and Accurate Deferral to Multiple Experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 154–165. https://doi.org/10.1145/3461702.3462516

[24] Agnes Mercedes Kloft, Robin Welsch, Thomas Kosch, and Steeven Villa. 2024. "AI enhances our performance, I have no doubt this one will do the same": The Placebo effect is robust to negative descriptions of AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–24. https://doi.org/10.1145/3613904.3642633

[25] Takanori Komatsu, Chenxi Xie, and Seiji Yamada. 2024. Waiting Time Perceptions for Faster Count-downs/ups Are More Sensitive Than Slower Ones: Experimental Investigation and Its Application. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3613904.3641942

[26] Chinmay Kulkarni, Stefania Druga, Minsuk Chang, Alex Fiannaca, Carrie Cai, and Michael Terry. 2023. A Word is Worth a Thousand Pictures: Prompts as AI Design Material. https://doi.org/10.48550/arXiv.2303.12647 arXiv:2303.12647 [cs].

[27] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[28] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 Publisher: SAGE Publications Inc.

[29] Haotian Li, Yun Wang, and Huamin Qu. 2024. Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642726

[30] Jenny T. Liang, Chenyang Yang, and Brad A. Myers. 2024. A Large-Scale Survey on the Usability of AI Programming Assistants: Successes and Challenges. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. ACM, Lisbon Portugal, 1–13. https://doi.org/10.1145/3597503.3608128

[31] David Chuan-En Lin and Nikolas Martelaro. 2024. Jigsaw: Supporting Designers to Prototype Multimodal Applications by Chaining AI Foundation Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3613904.3641920

[32] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3526113.3545621

[33] Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2023. 3DALL-E: Integrating Text-to-Image AI in 3D Design Workflows. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, Pittsburgh PA USA, 1955–1977. https://doi.org/10.1145/3563657.3596098

[34] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How AI Processing Delays Foster Creativity: Exploring Research Question Co-Creation with an LLM-based Agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–25. https://doi.org/10.1145/3613904.3642698

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

[36] Todd Lubart. 2005. How can computers be partners in the creative process: Classification and commentary on the Special Issue. *International Journal of Human-Computer Studies* 63, 4 (Oct. 2005), 365–369. https://doi.org/10.1016/j.ijhcs.2005.04.002

[37] M. Maher. 2012. Computational and Collective Creativity: Who's Being Creative?. In *International Conference on Innovative Computing and Cloud Computing*. https://www.semanticscholar.org/paper/Computational-and-Collective-Creativity%3A-Who%27s-Maher/dc07961727ea1b4ae2e66dfbe582750b9ab69a6e

[38] Joseph E. McGrath. 1995. Methodology Matters: Doing Research in the Behavioral and Social Sciences. In *Readings in Human–Computer Interaction*. Elsevier, 152–169. https://doi.org/10.1016/B978-0-08-051574-8.50019-4

[39] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3613904.3641936

[40] Jakob Nielsen. 1993. *Usability Engineering* (1st ed.). Morgan Kaufmann. http://gen.lib.rus.ec/book/index.php?md5=ce06f96ade58cd42e62424b0fa0d91e6

[41] Donald A. Norman. 1990. The 'Problem' with Automation: Inappropriate Feedback and Interaction, not 'Over-Automation'. *Philosophical Transactions of the Royal Society B: Biological Sciences* (May 1990). https://doi.org/10.1098/rstb.1990.0101

[42] Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference (Academic Mindtrek '22)*. Association for Computing Machinery, New York, NY, USA, 192–202. https://doi.org/10.1145/3569219.3569352

[43] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (June 1997), 230–253. https://doi.org/10.1518/001872097778543886 Publisher: SAGE Publications Inc.

[44] Raja Parasuraman and Christopher D. Wickens. 2008. Humans: still vital after all these years of automation. *Human Factors* 50, 3 (June 2008), 511–520. https://doi.org/10.1518/001872008X312198

[45] Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review*. Technical Report MSR-TR-2022-12. Microsoft. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[46] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv:2302.06590 [cs.SE] https://arxiv.org/abs/2302.06590v1

[47] Prolific. 2024. Prolific. https://www.prolific.com

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

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

[50] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290605.3300750

[51] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[52] Yingnan Shi and Bingjie Deng. 2024. Finding the sweet spot: Exploring the optimal communication delay for AI feedback tools. *Information Processing & Management* 61, 2 (March 2024), 103572. https://doi.org/10.1016/j.ipm.2023.103572

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

[54] Lucy Suchman. 2006. *Human-Machine Reconfigurations: Plans and Situated Actions* (2 ed.). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511808418

[55] Ulrik Söderström, Martin Bååth, and Thomas Mejtoft. 2018. The Users' Time Perception: The effect of various animation speeds on loading screens. In *Proceedings of the 36th European Conference on Cognitive Ergonomics (ECCE '18)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3232078.3232092

[56] Macy Takaffoli, Sijia Li, and Ville Mäkelä. 2024. Generative AI in User Experience Design and Research: How Do UX Practitioners, Teams, and Companies Use GenAI in Industry?. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1579–1593. https://doi.org/10.1145/3643834.3660720

[57] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–7. https://doi.org/10.1145/3491101.3519665

[58] Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2024. ReelFramer: Human-AI Co-Creation for News-to-Video Translation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20. https://doi.org/10.1145/3613904.3642868

[59] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581366

[60] Terry Winograd and Fernando Flores. 1987. On understanding computers and cognition: A new foundation for design: A response to the reviews. *Artificial Intelligence* 31, 2 (Feb. 1987), 250–261. https://doi.org/10.1016/0004-3702(87)90026-9

[61] Shishi Xiao, Liangwei Wang, Xiaojuan Ma, and Wei Zeng. 2024. TypeDance: Creating Semantic Typographic Logos from Image through Personalized Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–18. https://doi.org/10.1145/3613904.3642185

[62] Litao Yan, Alyssa Hwang, Zhiyuan Wu, and Andrew Head. 2024. Ivie: Lightweight Anchored Explanations of Just-Generated Code. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3613904.3642239

[63] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[64] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[65] Nur Yildirim, James McCann, and John Zimmerman. 2020. Digital Fabrication Tools at Work: Probing Professionals' Current Needs and Desired Futures. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376621

[66] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[67] Guanglu Zhang, Ayush Raina, Jonathan Cagan, and Christopher McComb. 2021. A cautionary tale about the impact of AI on human design teams. *Design Studies* 72 (Jan. 2021), 100990. https://doi.org/10.1016/j.destud.2021.100990

[68] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[69] Jiayi Zhou, Renzhong Li, Junxiu Tang, Tan Tang, Haotian Li, Weiwei Cui, and Yingcai Wu. 2024. Understanding Nonlinear Collaboration between Human and AI Agents: A Co-design Framework for Creative Design. (2024).

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

Received September 2024; revised December 2024; accepted January 2025