

Improving Experimental Methods to Capture Real-World Human-AI Perceptions and Interactions

Nikita Haduong

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Noah A. Smith, Chair
Steve Tanimoto
Yulia Tsvetkov

Program Authorized to Offer Degree:
Computer Science and Engineering

© Copyright 2025

Nikita Haduong

University of Washington

Abstract

Improving Experimental Methods to Capture Real-World Human-AI Perceptions and Interactions

Nikita Haduong

Chair of the Supervisory Committee:

Noah A. Smith

School of Computer Science and Engineering

AI agents are being increasingly used in production settings, but our understanding of how humans expect AI to behave, and how AI usage influences human behavior, falls short because of the gap between controlled laboratory studies and real-world usage. In this thesis, I develop methodologies to shrink this gap and further our understanding of how humans perceive and use AI in practice, and how we can design more relevant technologies. My methodologies are anchored by the observation that participants with greater task immersion and intrinsic motivation allow modeling more realistic behavior, and simple manipulation of task settings, domains, and incentives can increase immersion. This thesis discusses my key contributions to making AI research more relevant to potential downstream users.

I first consider the role of AI in collaborative problem solving (CPS) and discover a dearth of open resources for conducting research in human-AI CPS when teams are larger than dyads. I approach this challenge by developing `CPS-TaskForge`, a CPS environment generator based on a resource management task, hence resembling real-world problems. `CPS-TaskForge` enables systematic study of CPS and open data generation by parameterizing tower defense games, and is thus approachable to laypeople and intrinsically motivating because the task is *fun*. Next, I explore how potential risks and harms of AI assistants are perceived and understood by users by grounding the discussion in procedural document question answering which has tangible and relatable risks to human evaluators, and recruiting evaluators who are familiar with the domain of procedural documents. I discover how current human evaluation techniques fail to account for non-deterministic AI behavior and develop a taxonomy of errors that can help inform the future development of an AI-powered system. Finally, I examine AI-assisted decision making behavior and explore the influence of performance pressure, a common environmental factor

in production settings that lab studies isolate away from, to further our understanding of the sensitivity of AI advice taking. My methods illustrate the importance, and potential simplicity, of modeling more realistic deployment settings while conducting carefully controlled studies.

Acknowledgements

I am grateful to my advisor, Noah A. Smith, for believing in me and helping me grow, as both a researcher and a human, throughout this journey. Next, I would like to thank the members of my committee—Steve Tanimoto, Shane Steinert-Threlkeld, and Yulia Tsvetkov, for valuable insights and feedback that improved this thesis. I have been fortunate to collaborate with many brilliant minds—Elizabeth Clark, Tal August, Sofia Serrano, Sachin Gururangan, Alice Gao, Bo-Ru Lu, Chia-Hsuan Lee, Zeqiu Wu, Hao Cheng, Paul Koester, Jean Utke, Tao Yu, Mari Ostendorf, Chien-Yu Lin, Irene Wang, Prithviraj Ammanabrolu, Steve Seitz, and Zoran Popović, and enjoyed the camaraderie of Noah’s ARK, the UW NLP group, and GRAIL. I want to thank the Center for Statistics and the Social Sciences and Thomas Frye for providing valuable consulting services that improved the rigor of the works in this thesis. Finally, this would not have been possible without the support of my friends and family.

DEDICATION

To Texas, my family, friends, and all of our pets.

Contents

1	Introduction	17
1.1	Outline	18
2	CPS-TaskForge: Generating Collaborative Problem Solving Environments for Diverse Communication Tasks	21
2.1	Introduction	21
2.2	Collaboration and Problem Solving	23
2.3	CPS-TaskForge and Tower Defense	25
2.3.1	System Overview	27
2.3.2	CPS-✓ : A CPS Task Design Checklist	30
2.4	Case Study	32
2.4.1	Data Collection	33
2.4.2	Results	34
2.4.3	Observations	35
2.4.4	Analysis	36
2.5	Related Work	37
2.6	Limitations	38
2.7	Summary	39
3	Human-AI Partnerships: Risks and Harms in NLP	41
3.1	ProcDocQA	42
3.1.1	Risk-Aware Design Questionnaire	43
3.2	Case Study: ProcDocQA for Recipes	44
3.2.1	RADQ Initial Completion	45
3.2.2	User Perspective Study	46

3.2.3	Results	48
3.2.4	Improvement Results	49
3.2.5	MultiDecoding Analysis	51
3.2.6	RADQ Post-Study Update	55
3.3	Limitations	57
3.4	Ethics Statement	57
3.5	Summary	57
4	How Performance Pressure Influences AI-Assisted Decision Making	59
4.1	Introduction	59
4.2	Related Work	61
4.3	Experiment	63
4.3.1	Dataset	63
4.3.2	XAI Methods	64
4.3.3	User Interaction	64
4.3.4	Independent Variables.	65
4.3.5	Experiment Setup	65
4.4	Results	67
4.4.1	How did participants judge reviews?	70
4.5	Discussion	70
4.5.1	AI advice reliance under pressure	70
4.5.2	Practical Advice	72
4.6	Future Directions	72
4.7	Limitations	73
4.8	Ethical Considerations	73
4.9	Summary	73
5	Conclusion	75
5.1	Future Directions	76
A	CPS–TaskForge: Generating Collaborative Problem Solving Environments for Diverse Communication Tasks	101
A.1	Survey Questions	103

A.1.1	TWQ	107
A.1.2	VIA Team roles	108
A.2	CPS classification	110
A.2.1	Annotation challenges	111
A.2.2	Prompt	112
A.3	Potential CPS-TaskForge Tasks	114
A.4	License	115
B	Human-AI Partnerships: Risks and Harms in NLP	117
B.1	Surveys	120
B.2	Survey Results	130
C	How Performance Pressure Influences AI-Assisted Decision Making	131
C.1	Methods	131
C.1.1	RAIR and RSR	131
C.1.2	Presurvey	132
C.1.3	Postsurvey	133
C.2	Results	135
C.2.1	Significance Tests	135
C.2.2	Regression Tables	136
C.3	Licenses	136
C.4	AI Use	137

List of Figures

2.1	CPS-TaskForge in-game screenshot	26
2.2	CPS-TaskForge system overview	28
2.3	CPS-TaskForge pilot study level maps	35
2.4	CPS-TaskForge pilot study tower placement frequency	36
2.5	CPS-TaskForge pilot study player strategies	37
3.1	Dimensions characterizing potential harms of procedural document domains	42
3.2	Example annotation question with responses described in §3.2.4.	48
3.3	Distribution of the types of questions annotated	49
3.4	How annotators judged the correctness of machine and human responses	50
3.5	Output instability example	52
3.6	Leading question agreement, hallucination, and recommendation errors.	53
3.7	Hallucination and language style behavior	54
3.8	Multiple possible correct answers.	55
4.1	User interface in payment-LIME condition	66
4.2	Accuracy after receiving AI advice.	68
4.3	GenAI improves appropriate AI advice reliance, but pressure has a predominantly negative effect.	68
A.1	Money remaining for every team	103
A.2	Presurvey part 1	104
A.3	Presurvey part 2	105
A.4	Presurvey part 3	106
B.1	Presurvey part 1	120
B.2	Presurvey part 2	121

B.3 Presurvey part 3	122
B.4 Presurvey part 4	123
B.5 Postsurvey part 1	124
B.6 Postsurvey part 2	125
B.7 Screen shown when a response answers the question.	126
B.8 Screen shown when annotators state a response does and does not answer the question.	127
B.9 Screen shown when the response does not answer the questions.	128
B.10 Screen shown if the annotator is unsure if the response answers the question.	129

List of Tables

2.1	A sample survey of collaborative problem solving research	24
2.2	CPS-✓	30
2.3	Statistics of openly available corpora collected during a CPS task	34
2.4	Sample conversation from one team during Level 1, Round 2.	34
2.5	CPS Skill usage from our case study	35
3.1	Risk-Aware Design Questionnaire for ProcDocQA	43
3.2	Sample questions and answers from the recipes dataset.	44
3.3	Annotations for how responses could be improved.	49
3.4	Percentage of prompts for which each behavior was present in at least one of the ten responses generated	51
A.1	Example conversations and interactions from our CPS-TaskForge pilot study.	102
B.1	Example GPT-3 prompt(s).	118
B.2	Additional sample questions and answers from the recipes dataset.	119
C.1	Holt-Laury survey	134
C.2	Accuracy pre- and post- AI advice	135
C.3	GLMM results for RAIR	136
C.4	GLMM results for RSR	136
C.5	GLMM results for overall accuracy	136

Chapter 1

Introduction

Humans have worked towards developing artificial intelligence (AI) for decades, but only recently has the technology begun to show promise. In 2014, Microsoft released Xiaoice, a Chinese chatbot with great success: more than 40 million conversations were produced within a year (Ars Technica, 2016). Riding on that success, in 2016, they released the Twitter chatbot Tay to Western audiences. Despite implementing filtering and conducting “extensive user studies with diverse user groups”, Tay was taken offline within 24 hours due to producing offensive and inappropriate content (Lee, 2016). Users had unexpectedly attacked Tay.

A major turning point for AI in language occurred when OpenAI released the foundation model ChatGPT in 2022. This model could be accessed via an easy-to-use conversational web interface and simple API. Previously, access to advanced AI technology was reserved for technological experts. Now, with ChatGPT and other foundation models, anyone can interact with and release AI content with minimal knowledge of AI capabilities. In May 2023, a lawyer used ChatGPT to assist with research and submitted fabricated case documents, unaware that ChatGPT could hallucinate information (Neumeister, 2023). In February 2024, a customer service chatbot for Air Canada informed a customer they would receive bereavement compensation, but Air Canada denied their claim and was then sued. The customer won the case (Garcia, 2024). Both prior to and after the release of accessible foundation models, researchers across many communities—e.g., HCI (Cai et al., 2019), NLP (Bansal et al., 2019; Vats et al., 2024), and AI fairness (Lai et al., 2021; Tan et al., 2021; Jacobs and Wallach, 2021; Manheim and Kaplan, 2019; Raso et al., 2018), have been developing ways to effectively, responsibly, and safely integrate AI into human society. The companies releasing foundation models have also published research in the space: Anthropic is active in AI alignment (Bai et al., 2022; Radhakrishnan et al., 2023), Microsoft has several publications in AI safety (Rivera et al., 2020; Gausen et al., 2024), and Meta developed a

powerful and popular method to mitigate hallucination: Retrieval Augmented Generation (Lewis et al., 2020). Yet harms and negative consequences from using AI continue to make headlines.

Why are known undesirable AI behaviors still rampant in production, and how can we prevent negative outcomes from AI usage before the technology is deployed? In this thesis, I argue that the rapid adoption of AI outpaces our understanding of interactions between human and AI, resulting in a divide between how we study AI, in controlled lab settings, and how people use AI. To shrink this gap, I developed methodologies to strengthen the real-world relevance of research findings and provide actionable insights to AI application developers. My methods push the boundaries of past research by providing support for the modern technological landscape, i.e., where AI is rapidly being integrated and used by everyone without necessarily understanding the implications of using the technology. This thesis will discuss three settings: collaborative problem solving, risk assessment, and decision making, and how we can update our methodology to meet the needs of a world already abundant with AI.

1.1 Outline

An effective team can incorporate strengths from all members to produce a greater outcome than the sum of its parts (members). As a result, human teams, and understanding how to improve team performance, have been the subject of extensive research across many disciplines, from sports performance to cognitive science, psychology, and business organizations. Team activities are goal-oriented, and can be cast as a *collaborative problem solving* (CPS) activity, because the team members need to work together to reach their desired outcome, navigating both cognitive and social communicative challenges as a team, in addition to managing individual performance.

In Chapter 2, I explore the question of how we can design AI to support human teams. Much of our understanding of CPS communication is dominated by dyadic teams (human-human and human-AI), which is then reflected in open research resources, e.g., datasets (§2.1.); however, CPS teams often include more than two parties. Furthermore, CPS occurs under diverse environments and setups, such as with different team compositions, task characteristics, and problem scenarios, but current open resources are limited to the select few setups (e.g., pairs of participants, the search and rescue task, and the task of identifying objects) and are difficult to expand to alternate settings. *How can we develop AI tools to support CPS activity if we do not first know what support is required?* I approach this resource challenge by developing a CPS task environment generator, `CPS-TaskForge`, that enables systematic study of CPS through the parameterization of team composition, task characteristics, and problem scenarios. To

ensure task realism, `CPS-TaskForge` is based on a resource management task—a common real-world CPS activity (e.g., scheduling a meeting). I also release `CPS-✓`, a CPS task design checklist adapted from the PISA 2015 CPS Framework (OECD, 2017), that can be used as a standalone tool for designing CPS research or in conjunction with `CPS-TaskForge`. To validate `CPS-TaskForge`, I conduct a case study and show how the data generated is diverse and can be used in further CPS research.

Technology can be harmful when it acts in unintended ways or users behave unexpectedly. Modern large language model (LM)-based AI tools can generate believable content with an authoritative tone, thus convincing humans to trust the output despite the fact that these tools are simply probabilistic models that can make up content, regardless of any instructions to produce only factual output. During the product development process, a developer will conduct risk analysis and design mitigations for known risks before deploying to users. The literature around risks of AI has often been theoretical and hypothetical, because the studies were conducted before current advances in LM-based AI, making it difficult for a developer to put into practice AI risk mitigations.

In Chapter 3, I design a method to assist with developing more risk- and harm-aware AI systems by first grounding these theoretical discussion in concrete applications. I then release a questionnaire that can help guide developers to produce safer applications and create a novel method of analyzing LM outputs that reveals insights into model behavior. I illustrate the application of my methods through a case study and show how some behaviors seen as harmful in theory can actually be desirable for some users or domains.

Decision making is an important activity required in work, with a range of associated stakes, from minimal to life-dependent. When making a decision, a human may seek advice, particularly if they are not confident in their decision. This advice can come from a human or a machine, and the social dynamics and personality of the decision-maker play a role in how they receive the advice. For example, if a human receives advice from someone they deem inferior, they may ignore the advice regardless of its correctness. If they receive the same advice where the stakes are high, they may behave differently. In the case of advice from a machine—AI, humans have biases and preconceptions about machine capabilities that can influence how they use the advice. As a result, when a human collaborates with an AI assistant that has objectively higher task performance than the human, the overall joint performance can still fall short (Smith, 2021; Busuioc, 2021; Johnson, 2021; Schemmer et al., 2023). How do environmental factors such as pressure influence how humans take advice from an AI assistant?

In Chapter 4, I explore the sensitive nature of AI-assisted decision making by asking humans to make decisions under performance and time pressure. My results paint a complex picture of the powerful

effect of pressure, where pressure can both improve and worsen advice-taking behavior, overriding benefits from AI-assistive tools already carefully designed to mitigate the negative biases of humans against AI advisors. I show the importance of accounting for the effects of pressure when designing systems intended for use in environments with pressure. I illustrate the ease of applying different types of pressure to studies by manipulating monetary and temporal performance pressure and argue for the importance of modeling realistic deployment environments in human-AI research.

My contributions aim to improve the capture of realistic human-AI collaborative behavior in experimental research. I developed methods applicable to different stages of the research process, showing how improvements can be made at any point, from data curation to simulating deployment settings. I encourage researchers in human-AI applications to be cognizant of how their work would be received and transformed by society, and to conduct preventative error analysis based on relevant real-world factors.

Chapter 2

CPS-TaskForge: Generating Collaborative Problem Solving Environments for Diverse Communication Tasks

Collaborative problem solving (CPS) is a skill required to succeed in modern life that is challenging to develop, requiring both cognitive and social abilities. Including AI agents in a team can potentially increase team effectiveness, but before studying how AI agents can effectively be integrated to support teams, we must first understand how humans behave in teams and where they would benefit from AI assistance. Unfortunately, open research resources primarily focus on dyads (two humans, or a human and an AI), resulting in barriers to research in CPS with larger teams. This chapter describes my approach to resolving this barrier through the development of a CPS task environment generator, CPS-TaskForge. CPS-TaskForge can generate task environments with close resemblance to real-world tasks and supports multiple team compositions, thus paving the way to future research in CPS with diverse human-AI teams. Work in this chapter was published as Haduong et al. (2024), and all code and data is released at <https://github.com/nhaduong/cps-taskforge>.

2.1 Introduction

Modern life requires teamwork to solve problems (Marks et al., 2001), but what makes a team work well together? This area of study, known as collaborative problem solving (CPS), is active across many

disciplines, e.g., psychologists study the construction of team mental models in team discussions Lee (2015), business management sciences investigate how communication style affects performance evaluation (Proell et al., 2022), and educators develop tools to teach team communication strategies (Stewart et al., 2023), emphasizing the research direction of discovering *how team members talk to one another*. Conducting empirical work in CPS faces many challenges, in large part because of a large CPS task design space (e.g., what is the problem, who makes up the team, and who knows what information when). As a result, despite extensive interdisciplinary work in CPS, task designs in empirical studies have often focused on teams of two collaborating to solve problems such as selecting a designated object, modeling search and rescue, and making decisions.

AI agents have the potential to increase team effectiveness, and developing ways to integrate AI into teams is an active area of research in communities such as HCI (Cai et al., 2019), NLP Bansal et al. (2019); Vats et al. (2024), and AI fairness (Lai et al., 2021). Example integrations include AI-assisted decision making with one human and one AI (e.g., cancer diagnosis, Chen et al., 2021) and AI-assisted creative tooling (e.g., Tsiros and Palladini, 2020; Lu et al., 2024a). Developing these collaborative tools is made possible through open datasets. For example, various Amazon reviews datasets (e.g., Fornaciari and Poesio, 2014 and Ni et al., 2019) have been used to develop sentiment classifiers and deception detectors that can be used as AI-assisted decision makers, and the Reddit WritingPrompts dataset Fan et al. (2018) has been valuable in developing co-writing AI systems. Unfortunately, a paucity of open datasets with more than two parties leads to challenges in integrating AI with larger human teams, as we lack understanding of team dynamics when an AI communicates to a team, rather than an individual.

To support CPS study across different designs (e.g., adding a third AI teammate to a two-human team or using voice instead of text communication), we introduce a CPS task environment generator, `CPS-TaskForge`. `CPS-TaskForge` instantiates a *resource management* activity through a **tower defense** game and supports adjusting a range of CPS design parameters such as team composition, communication method, and how stressful the task is. In a tower defense game, players must defend their base by using limited resources to construct towers that can defeat enemies before the enemies destroy the base. We provide a CPS task design checklist, `CPS-✓`, adapted from the PISA 2015 theoretical CPS framework (PISA2015) developed by the OECD (OECD, 2017), to support generating the desired task environment with `CPS-TaskForge`.

We illustrate `CPS-TaskForge` capabilities by presenting several CPS task designs and conducting a case study that can collect human communication data exhibiting a range of CPS skills, including social skills such as maintaining group communication and cognitive skills such as developing strategic

plans. Our study has small groups of 3–4 participants complete a task multiple times with increasing difficulty. We observe many different successful strategies and a wide range in CPS skill usage across teams, demonstrating the versatility of collecting data through CPS-TaskForge.

To summarize our contributions:

1. We identify opportunities and gaps in the interdisciplinary CPS literature. We argue that human team research can help advance human-AI team design; however, there exist challenges associated with the lack of diverse CPS data available to the research community.
2. We introduce CPS-TaskForge, which allows researchers to generate a variety of CPS task environments for studying human and human-AI CPS team processes. We adapt a theoretical CPS framework into a design checklist, CPS-✓, to assist with CPS-TaskForge environment generation.
3. We present a case study using CPS-TaskForge to illustrate the variability of CPS data through a study with more than two agents. We release the conversation and game interaction data collected during the study as an example of what can be produced using CPS-TaskForge.

2.2 Collaboration and Problem Solving

Collaborative problem solving (CPS) processes are well-studied for human teams, but when human-AI teams are considered, downstream task performance has been prioritized, leaving human-AI CPS processes understudied. For example, Proell et al. (2022) found human team communication more effective when the appropriate style was used in conjunction with the delivery of relevant information. Humans have different expectations towards AI teammates (Zhang et al., 2023, 2021; Grimes et al., 2021), so human-AI teams may value communication style differently. Studying human-AI CPS processes requires developing the appropriate datasets, but resources for creating such data is deficient.

Understanding how effective and efficient communication can predict successful teamwork requires collecting data in a variety of CPS settings. The tasks used to elicit relevant data often model real-world activities, e.g., rescuing humans from a burning building (ASIST; Corral et al., 2021; Freeman et al., 2021), instruction following through selecting designated objects (e.g., PentoRef, Zarriß et al., 2016; KTH Tangrams, Shore et al., 2018; PhotoBook, Takmaz et al., 2020; Doll Dialogue, Tenbrink et al., 2017; Paxton et al., 2021), and navigating environments (e.g., HCRC Map Task, Anderson et al., 1991; Effenberger et al., 2021), and use human participants. The resulting datasets have been used to study a

	Task type	Team Size	Communication Modality
KTH Tangrams (Shore et al., 2018)	Object Identification	2	Speech
PentoRef (Zarrieß et al., 2016)	Object Identification	2	Multimodal
TEAMS (Rockenbach et al., 2007)	Forbidden Island TM	3–4	Multimodal
ASIST (Huang et al., 2022)	Search and Rescue	3	Multimodal
CerealBar (Suhr et al., 2019)	Search and Rescue	2	Text
HCRC Map Task (Anderson et al., 1991)	Search and Rescue	2	Speech
PhotoBook (Takmaz et al., 2020)	Object Identification	2	Text
Cards (Potts, 2012)	Search and Rescue	2	Text
Rodrigues et al. (2021)	Object Identification	2	Multimodal
Ma et al. (2023)	Programming	2	Multimodal
Butchibabu et al. (2016)	Search and Deliver	2	Text
Kokel et al. (2022)	Object Construction	2	Multimodal
• MRE (Hill et al., 2003)	Decision Making	21	Speech
T-shirt Task Andrews et al. (2019)	Math Problem	2	Multimodal
Volcano Lab Flor et al. (2016)	Science Lab	2	Text
Circuit Lab Graesser et al. (2018)	Science Lab	3	Text
Physics Playground (Sun et al., 2020)	2D Physics Puzzles	3	Multimodal
Minecraft (Sun et al., 2020)	Minecraft Hour of Code	3	Multimodal
CPSCoach (Stewart et al., 2023)	2D Physics Puzzles	2	Multimodal
• NeoCities Schelble et al. (2022)	Search and Rescue	3	Text
9-11 Firefighting (Hutchins et al., 2008)	Firefighting	—	Speech
Air Warfare (Hutchins et al., 2008)	Object Identification	6+	Speech
MIO* (Hutchins et al., 2008)	Object identification	3+	Speech
Wiltshire et al. (2018)	NASA Moonbase Alpha Simulation	2	Speech
CPS-TaskForge (this chapter)	Object Identification, Resource Management	1–4+	Text, Speech

Table 2.1: A sample of collaborative problem solving research. MIO* is short for Maritime Interdiction Operations. The top group contains work that produced datasets open to the research community. • indicates studies with AI teammates. Object identification tasks require identifying an object, search and rescue requires navigating an environment to locate an object, and search and deliver requires returning to a second point after locating the object. The math and science lab tasks are typical tasks found in educational contexts. Forbidden IslandTM is a commercial cooperative board game. “Text” data often contains system interaction log data such as mouse clicks, whereas “Multimodal” communication may include video of participant bodies, audio, and hormonal measurements. We observe more diverse tasks conducted in works without open data.

wide variety of communication and linguistic phenomena, including language entrainment (i.e., when communicative behavior becomes similar among interlocutors, including lexical choice and rhythm) and common ground building (i.e., when interlocutors develop their own code). To the best of our knowledge, analogous settings incorporating an AI team member in a CPS task have not explored similar communication and linguistic phenomena because only recently has AI-generated natural language become indistinguishable from humans (Clark et al., 2021a; Dugan et al., 2022), enabling exploration of AI teammates as peers. Unfortunately, expanding pre-existing datasets to other CPS settings, such as involving an AI agent or a third human team member, is challenging because the tasks were designed to study a specific team composition; for example, what role would a third participant play in a navigation task originally designed for one human to tell another human where to go?

Despite the extensive body of literature studying CPS, publicly available resources remain scarce, particularly when more than two agents are involved. We summarize a sample of CPS task activities in the literature in Table 2.1 to illustrate gaps in task type and team size between studies with or without data release to the research community.

2.3 CPS-TaskForge and Tower Defense

To advance CPS research, we need ways to systematically study CPS when varying factors, allowing comparison of CPS results across settings. We therefore develop a CPS task environment generator, CPS-TaskForge, which can generate CPS environments with different design factors. We also release a CPS task design checklist, CPS-✓, that describes how varying design factors produces different environments. We defer discussion of CPS-✓ to Section 2.3.2; here we give a concrete description of the task environments our work targets.

We start with several requirements: (R1) CPS-TaskForge should be built on an activity that can support the different values in CPS-✓; (R2) the activity should be **fun**, to motivate participant signups, because CPS studies require multiple participants, making scheduling a logistical barrier to conducting CPS research; (R3) the activity should be easy to learn for both participants and researchers, in order to minimize time spent in tutorials and allow researchers to quickly design different CPS studies; and (R4) the activity should easily scale in difficulty to enable CPS research studying effects of expertise on collaboration.

We meet our design requirements by using the Tower Defense (TD) game genre as our CPS-TaskForge activity. The premise of a TD game is to defend a base from enemies by placing tow-



Figure 2.1: In-game screenshot of a game produced by CPS-TaskForge, used in our case study. Enemies spawn from (1) and can only move on the brown path. Towers can only be placed on the green spaces. (2) is the timer used during the *planning* phase, indicating how much time players have to set the board before the *attack* phase starts. (3) tracks base health—players lose if it drops to zero due to enemies reaching the base, the amount of money available to purchase towers and upgrades, and a running score. (4) is the set of towers this player can build. Different towers have different abilities and costs. (5) previews the enemy sequence of a spawn point. (6) is the text chat players use to communicate with each other. (7) is the base players must defend. (8) is an upgrade menu for a selected tower. (9) is an information panel about a tower. A coordinate grid is provided so players can refer to specific spaces on the map when communicating with each other.

ers on the map, which can destroy the enemies. TD games require strategy and resource management—a vital aspect of CPS tasks Care et al. (2015)—and games have been successfully used by the research community to study communication (e.g., Codenames, Shaikh et al., 2023) and collect data (e.g., Verbosity (von Ahn et al., 2006), Duolingo (von Ahn, 2013), SearchWar (Law et al., 2009), and MatchIn, Hacker and von Ahn, 2009).

TD games are known for having a gentle learning curve, short levels (R3), and ease in scaling difficulty through simple designs (R1, R4; Avery et al., 2011). The 2021 mobile market value for TD games was estimated at 940 million USD Analytica (2022); this popularity suggests the potential for participants to play the game of their own volition (R2). It is also known to support 1–4 players in cooperative play,¹ natively supporting studying human-AI teams involving as few as one human.

We briefly describe what a TD game involves, referencing an in-game screenshot (Figure 2.1) of an environment produced by CPS-TaskForge. In a TD game, the player needs to defend their base (7) from enemies by placing towers (9) on the map whose inhabitants can attack the oncoming enemies. The

¹Bloons TD 6™ is a commercial game with a 4-player cooperative mode.

enemies will appear at designated spawn points (1) and traverse the map along specific paths known to the player, allowing the player to strategize where to place towers effectively. Players must manage their resources (3) (e.g., gold and map real estate) when developing their defense strategy. Levels differ in the enemy spawning behavior (e.g., enemies can spawn without a break, or there is time in between groups of enemies), enemy variants (e.g., a faster or slower enemy), map terrain (e.g., obstacles can prevent tower placement), and player resources (e.g., types of towers, amount of starting gold). The standard TD game has two phases: *planning*, a static phase where players can place towers on the map, and *attack*, a dynamic phase during which enemies spawn, and players can react to the changing situation by adjusting their towers.

2.3.1 System Overview

CPS-TaskForge is built using the open-source game engine Godot,² Nakama,³ and data collection uses REST API calls to an external server.⁴ All code within Godot is written in GDScript. Godot has native support for multiplayer networking, text localization, and game design content can be saved to human-readable text-based formats, allowing researchers to design environments with minimal knowledge of Godot. It also has an active plugin ecosystem that enables easy extensibility, including AI agent plugins (e.g., Godot RL Agents (Beeching et al., 2021) and GodotAgent⁵) for conducting human-AI research. Multiplayer syncing and logic is handled server-side, e.g., the server communicates the game state to clients, rather than game logic being computed on the client, and the client communicating to all other clients the updated game state. For example, suppose a client player wants to upgrade a tower. The player interacts with the upgrade button, which sends a purchase request to the server. The server determines if the purchase is permissible, then communicates to all clients the new game state (an upgraded tower, if the purchase was permitted). Player game interactions (e.g., purchasing, upgrading, and selling a tower), communication, and game scores are logged to the external server by default. Additional data logging can be added as needed. CPS-TaskForge supports moderated sessions, where the researcher can enter the game to observe gameplay without acting as a player, and unmoderated play, where players can run sessions on their own. The game host is designated as the server for multiplayer, and a client player can simultaneously be the server.

²<https://godotengine.org>

³<https://heroiclabs.com/nakama/>

⁴The external server we release alongside CPS-TaskForge is a Python Flask server.

⁵<https://github.com/Wizzerdd/GodotAgent>

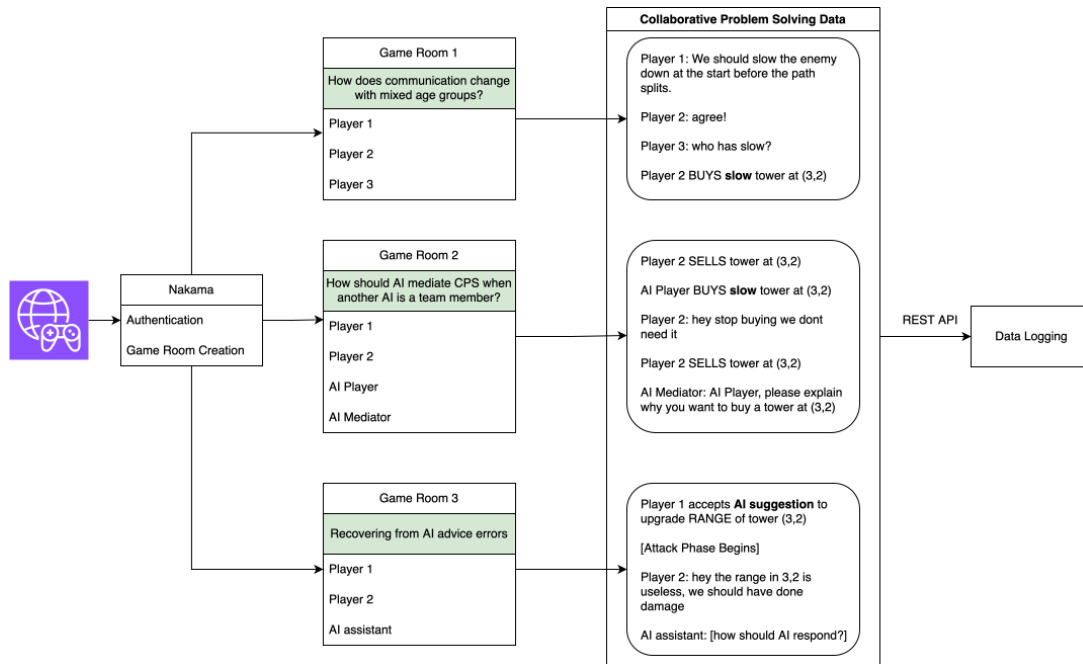


Figure 2.2: System overview illustrating 3 different research questions that CPS-TaskForge supports. Players authenticate through Nakama, join game sessions with different experimental environment designs driven by research questions, and generate CPS data while playing the game. Player interactions and communication are collected using REST APIs.

User Experience The experience flow is depicted in Figure 2.2, which we describe here. First, the game executable is distributed to all players. Players authenticate through Nakama, then either a player or the experimenter (in a moderated session) hosts a game room. The host distributes the unique room key generated by Nakama to all other players. Players join the room and see a random team name that they can edit. The purpose of the team name is to improve team cohesion and collaboration through the construction of a group identity (Carron and Spink, 1993). After all players have joined the room, the host starts the game. Players then play levels as designed by the experimenter (e.g., one level or multiple rounds per level). At the end of a round, a leaderboard is displayed with the team name and score breakdown. Leaderboards are known to improve user performance (Mekler et al., 2013; Landers et al., 2017), and they allow teams to track their progress against themselves (for tasks with multiple rounds per level) and others.

User Interaction. Each player is given a unique color that is used in the text chat display. The color is also used to outline the towers they placed (Figure 2.1; purple color) to indicate who placed which tower. Towers can be placed by clicking a button (Figure 2.1; 4) or through the assigned hotkey. Tower information is shown in a panel (Figure 2.1; 9) that appears when any tower is targeted. Selecting a tower will open an upgrade panel. Upgrades are given extra visual effects to help players understand

the game state and mechanics (Zhou and Forbes, 2022): upgrading the range that a tower can interact with alters the size of a colored circle around the tower, damage upgrades are indicated by the quantity of sparkles surrounding a tower, and firerate is shown through the speed of the orbiting sparkles. The addition of visual effects gives players an idea of which upgrades are applied to towers without needing to target towers to open the information panel.

CPS Interface Designs. To facilitate CPS communication behavior, we include several user interface design parameters not commonly found in TD games that can be toggled and customized as needed. Tower names can be hidden, which creates a setting similar to those used in common ground building studies, as players will need to develop a code to refer to specific towers. We provide a preview of the sequence of oncoming enemies from a spawn point (Figure 2.1; 5), which is vital to experiments conducted without the dynamic attack phase. The preview gives information that players can use to plan their strategy, and enables longer level designs without requiring players to memorize the enemy spawn behavior if players can play a level multiple times. We provide a coordinate grid label across the map so that players can refer to specific locations, in a similar manner to chess coordinates. Features can be disabled depending on the experimenter's study goal, e.g., if the research goal is to investigate how different teams refer to a particular location, the experimenter may want to disable the coordinate grid label.

Tower Defense Designs. Currently implemented tower defense designs that can be adjusted to suit the specified CPS task are as follows.

- Communication modality: Voice chat (bool), push-to-talk (bool), text chat (bool)
- Description visibility: Tower name (bool), tower description (bool)
- Number of rounds per level (int)
- Player resources: Money (shared, bool), health and score (shared, bool)
- Interactability during attack phase (bool). If enabled, then players can adjust tower placement and upgrades during the dynamic attack phase.
- Towers: We provide 12 custom towers with unique mechanics and effects. Information about towers (name, description) can be customized. The unique towers are: basic, poison (damage over time), piercing (damage multiple enemies in a straight line), splash (area damage), obstacle (spawn

an object on the track that does damage when enemies walk over it), slow (slows enemies), fear (enemies go backwards along the track), sniper (does more damage to faster enemies), discount (lowers upgrade costs of nearby towers), support (buffs all stats for nearby towers), and multishot (shoots in 4 directions).

- **Levels:** A level design designates how enemies spawn, the enemy movement paths, the location of a base that players defend, terrain for where towers can be placed, starting gold and health, and which towers are available to players.
- **Enemies:** There are enemy variants that differ in health, movement speed, point value when destroyed, and money given to players when destroyed.

What are we studying? E.g., Decision making, collaborative learning, negotiation, communication under stress		
Context	Dimension	Example Values
Problem Scenario	Q1. How is the task evaluated for success?	Binary win/lose, score(time, health)
	*Q2. How long does one CPS instance take to complete?	1 minute for planning and 1 minute for attack
	*Q3. How do skill and expertise scale with repetition?	Levels of similar difficulty are repeated, level difficulty scales by introducing more enemy spawn points
Team composition	*Q4. What fraction of teammates are human or AI?	H-H-H, H-AI, H-AI-AI, H-H-AI
	Q5. What is the symmetry of roles?	2 players have the same support towers, and 1 has all offense towers
	Q6. How are teammates interdependent?	Support towers are necessary to beat the level
Task characteristics	Q7. How open is the solution space?	Only 1 tower placement configuration can win
	Q8. What information is available, and how is new information distributed (if applicable)?	All players have the same information at all times, players must discover enemy spawn sequence
	Q9. How much stress are players under?	No stress (unlimited planning time)
Medium	Q10. What is the communication medium?	Text, voice

Table 2.2: CPS- \checkmark : Design questions adapted from PISA 2015 CPS design contexts. Questions with * are added to help design studies where task repetition is a dependent variable or considerations for human-AI teams. H = human.

2.3.2 CPS- \checkmark : A CPS Task Design Checklist

The PISA 2015 CPS Framework (PISA2015) (OECD, 2017) describes CPS tasks through a set of 15 design factors, showing how different CPS settings can be studied by manipulating different combinations of factors (e.g., team size and composition). To operationalize CPS research goals as design parameters that CPS-TaskForge can use to generate the environment, we define CPS- \checkmark , a design checklist adapted from PISA2015 (Table 2.2). We provide default values for CPS- \checkmark items in the event that some items are unnecessary to adjust for a particular study. We next explore how different hypothetical research goals can be targeted with different TD games generated by CPS-TaskForge and designed with the help of completing CPS- \checkmark .

Goal: Compare solution quality between all-human teams and mixed human-AI teams. To compare solution quality, we require a more complex task evaluation function than a simple binary win/lose value (Q1). We can design a scoring function to incorporate the time required to agree on a strategy during the planning phase, the amount of money used, or the distance enemies travel. We can also adjust the solution space size (Q7). A level can have a single solution, requiring a specific strategy for placing towers, and solution quality is evaluated by the speed of figuring out the solution. A level can also have multiple solutions, with solutions rated for quality, e.g., a solution using the minimum amount of towers is harder to achieve than a solution maximizing resource consumption and is thus higher quality. The solution quality comparison between teams can then measure the rate of solving levels with minimal resource consumption.

We want to use team compositions with different fractions of human and AI players (Q4). We can investigate how different team roles and personalities in all-human or mixed human-AI teams affect solution quality (Q5); for example, an all-human team where everyone identifies as a leader and has the same towers could result in poor solution quality due to an increase in conflict over strategy; or a team where a human leader effectively uses support towers from an AI teammate (Q6) may outperform a team with an AI leader who does not request support towers from a human teammate. Since we are interested in manipulating team composition, we can give all players a shared resource pool so that information is updated and distributed to all players simultaneously (Q8).

Goal: Investigate how stress affects team performance and communication. Stress can affect team performance, learning, and communication (Pfaff, 2012; Savelsbergh et al., 2012; Orasanu et al., 2004), with more successful teams developing adaptive strategies Kontogiannis and Kossiavelou (1999). We can model stressful situations by adjusting the amount of starting resources (money and planning time) to require more dynamic gameplay during the attack phase, forcing players to adapt to a rapidly changing environment (Q9). To design levels requiring more dynamic gameplay, we limit the initial starting resources such that players cannot beat a level by only placing towers during the planning phase. As enemies are defeated, players gain additional gold to spend towards placing more towers and upgrading existing towers, which are required to successfully defend their base. The control condition can then be giving players plentiful starting resources. We will evaluate the task with a simple binary win/lose (Q1) and allow several possible solutions so that teams are not discouraged if they cannot land on the single most optimal solution (Q7). Giving less money and planning time means players have to monitor the changing situation during the attack phase. We enable voice communication (Q11) so that typing speed

is not a factor.

Goal: Reimplement and extend prior work. Although `CPS-TaskForge` is designed to generate TD games, we can simulate object selection and manipulation tasks by limiting player interaction.

Object Selection. Reference games used in `KTHTangrams` (Shore et al., 2018) and `PentoRefTake` (Zarrieß et al., 2016) are played with two players in the roles Instruction Giver (IG) and Instruction Follower (IF). Both players have a view of the map. The IG is given the game goal (select a specific piece), and the IF can manipulate the map (select the piece). We simulate this task using `CPS-TaskForge`, by designing levels with towers placed on the board at the start, replacing the tower imagery with a pentomino or tangram. We enable voice communication and end the level upon a single tower object selection, evaluating success through whether the correct tower was selected (Q1).

Object Manipulation. Tenbrink et al. (2017) designed a task for furnishing a physical dollhouse. The IG is given the furnished dollhouse, and the IF is given an empty house. The IG needs to instruct the IF to furnish the house, and task success is evaluated by the correctness of object location and orientation. To simulate this task in `CPS-TaskForge`, we design levels that resemble house interiors, with walls designating rooms and preventing towers from being placed on them. We give the IF a set of towers that can be placed in the level, replacing the tower imagery with furniture. A tower can span multiple grid spaces on the map, and there are multiple copies of each tower with different orientations. The IG is provided the same level but with towers placed on the map already (similar to the setup for the reference games). Voice chat is enabled for communication. Since `CPS-TaskForge` produces digital grid-based games, object location and orientation can be automatically evaluated for correctness, improving upon the original setting, where evaluation was manually coded. A limitation of our simulation is that the original task used a physical dollhouse, giving participants multiple perspectives of the board (which could increase task complexity), while our simulation only gives players a single top-down view. 3D simulations or creating multiple 2D perspectives could be explored in future work.

2.4 Case Study: Communication of Small Groups as Task Difficulty Increases

To validate its flexibility, we want to explore whether `CPS-TaskForge` is capable of producing an environment that elicits diverse collaborative problem solving behavior. Prior work in CPS primarily used tasks with dyads or task repetitions at the same difficulty level, so we design a CPS task where

teams of 3–4 people complete a task, aiming to minimize expenditure of gold, at multiple difficulty levels.

We design our CPS-TaskForge environment as follows, referencing the questions from CPS-√. Task success is evaluated by the amount of money left unused, enemies destroyed, and health of the base (Q1). A single level takes 5–8 minutes to complete, depending on level difficulty, and we design 3 levels with increasing difficulty (Figure 2.3; CPS-√ Q2–3). All players are human (Q4), and each player is given 2–4 unique towers from a pool of 12 towers with different properties (§2.3.1) so that players have different roles, encouraging all players to engage and suggest usage of their own towers (Q5–6). Players are provided a surplus of gold, and costs are balanced to slightly favor upgrading over placing more towers, giving teams the opportunity to find many successful strategies (Q7). All new information is distributed to players simultaneously (e.g., how much damage an enemy receives from a tower) (Q8). Players are under moderate time stress because each level is calibrated to give ample but limited time (5–6 minutes) to discuss strategy and place towers, and we disabled interaction during the attack phase (Q9). Players could end the planning phase early. We designated level-specific planning time to ensure the study is completed in a reasonable amount of time. Players can only communicate through text chat (Q10). These design decisions showcase the simplicity with which the TD genre affords the ability to create different CPS task environments.

2.4.1 Data Collection

12 teams of 3–4 people (total 42 individuals) were recruited to participate in a 1.5-hour study⁶ and compensated with a gift card at a rate of 20 USD/hour. The study was conducted both in-person and remotely, and all studies were moderated. Recruitment occurred through school email listings and paper flyers posted around town. Participants were aged 18–24 (72%), 25–31 (18%), and 32+ (10%); 55% of participants were current undergraduates and 36% were in a graduate degree program; a third of participants rated their tower defense game familiarity below 3 on a 5-point Likert scale. Familiarity between teammates was not controlled, allowing some team compositions to contain strangers and others a subset of friends.

The study began with individual pre-surveys collecting basic demographic information, then participants watched a tutorial video explaining how to play the game and played a simple tutorial level together to become familiar with the interface. After the tutorial, they were given time to ask any questions about how to play the game. They then played 3 different levels 3 times each for a total of 9 games.

⁶Our local IRB approved our study.

Subsequent levels increased in difficulty, but the three rounds were the same for each level. Finally, they completed individual post-surveys containing questions about teamwork quality, team role identity, and team communication.

2.4.2 Results

	Teams	Participants	Team Size	Tokens	Size	Repetitions	Round Dur.	Study Dur.	Recruitment
TEAMS	63	252	3–4	573k	110k utts.	2	30min	1.5hr	Local
ASIST	64	192	3	—	—	2	15min	3.5hrs	Online, Local
CerealBar	N/A	264	2	325k	24k utts.	N/A	16.5min	—	Crowdworker
PhotoBook	N/A	1,514	2	984k	164.6k utts.	N/A	—	14.2m	Crowdworker
HCRC map task	32	64	2	150k	18hrs	4	—	—	School
PentoRef	63	127	2	216.3k	23k utts.	—	—	—	—
KHTTangrams	42	84	2	68k	11hrs/15k utts.	—	—	15min	Local
Cards	N/A	—	2	282k	45,805 utts.	N/A	8.5min	—	Crowdworker
Our case study	7	31	3–4	8k	1.5k utts.	9	4-6min	1.5hr	Local

Table 2.3: Statistics of openly available corpora collected during a CPS task. Repetitions are the number of tasks rounds completed by each team. Study durations are often longer than the time required to complete each round because they include surveys. Local recruitment indicates the local community and can include members beyond the research institution. — indicates information was not reported. Datasets with crowdworkers did not control for the number of repetitions workers could complete, and teams did not necessarily have unique workers, therefore stats reported are N/A.

Mundert: so im not placing my catapults if we do that
oobma: how many cannons then
oobma: 4 more?
omar: maybe 2?
Mundert: sure
Mundert: hoewver we can afford and more trees and vines too right
TommyVCT: wait
TommyVCT: should I sell my diamonds?
Mundert: maybe those crossbow things in the line as well
Mundert: not all
Mundert: right
Mundert: because slow is also good
omar: sell the diamonds in tile (8,9) and (8,8)
oobma: imo the cross bows would be good at 8,9
oobma: and 8,8
omar: ill putt a cross bw there
Mundert: agree
TommyVCT: That's all I got
Mundert: >
Mundert: ?
TommyVCT: The tank or controller like thingy is for faster emenies
Mundert: wait why is the tank there
omar: but could you sell tile 8,9?
TommyVCT: oh I put there
omar: crossbow is better there
Mundert: agree
Mundert: aight
Mundert: nice
omar: much better
Mundert: i dont think we need the tank
TommyVCT: yeah it's kinda useless
Mundert: more tree and vine and other such area of affect towers

Table 2.4: Sample conversation from one team during Level 1, Round 2.

We logged data using XML tags, and the data logged was text communication, score, and tower

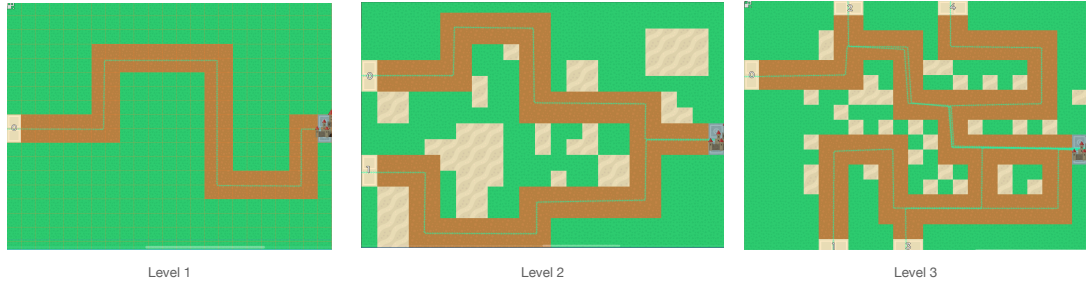


Figure 2.3: Level maps used in the CPS-TaskForge case study. Players can only place towers on the green spaces. Enemies spawn at labeled spawn points and move along the brown paths to the castle on the right. Level difficulty was scaled by introducing more enemy spawn points and limiting the green spaces for tower placement.

Dimension	CPS Skill	Example	Count	Avg. Tokens
Social	Maintaining communication	“haha okay”	222	2.3
	Sharing information	“I have a tower damage all enemies”	114	7.0
	Establishing shared understanding	“what does the diamond tower do?”	67	5.4
	Negotiating	“do we want to risk getting rid of anything else?”	38	5.4
Cognitive	Representing and formulating	“fires in multiple directions”	105	9.3
	Planning	“ok we can chokepoint the corners”	227	7.2
	Executing Actions	“k i maxed [upgrades]”	42	5.9
	Monitoring	“50 seconds D:”	86	5.0

Table 2.5: CPS Skill usage from our case study. Descriptive statistics are from the human annotated data (30% of the full dataset). Utterances were tokenized using the Spacy en_core_web_sm model.

interaction (upgrading, placing, and selling). The metadata associated with the data was the coordinates of interacted towers, timestamps, and the user. The first 4 teams were used to calibrate game difficulty and level designs and the data from one team was excluded from analysis because a team member left early, resulting in a final dataset of 7 teams producing 1.5k utterances with a vocabulary size of 1.2k (Table 2.3).

2.4.3 Observations

We adapt a CPS skill taxonomy developed by Andrews et al. (2019) to describe the communication data, simplifying the initial 10 skill taxonomy to 8 because of low annotation reliability (Table 2.5).⁷ We label only explicit natural language communications—the original taxonomy also includes system interactions (e.g., the act of placing a tower could be classified as “executing action”). A sample of 45 utterances of the data was manually annotated by two authors (inter-annotator agreement of 73%), then one author annotated 3 games (30% of the data). Example team communication is in Table 2.4 and Appendix Table A.1, exemplifying planning and directing through natural language, as well as communication through game behavior (e.g., placing a tower at a specified location when requested

⁷We discuss annotation challenges in Appendix A.2.1.

without using language to acknowledge the request.)

Cognitive CPS skills were used 49% of the time, and 29% of all communication was devoted to developing strategic plans (planning and negotiation skills). Andrews et al. (2019) observed 30% cognitive skill usage using a traditional collaborative math task, suggesting that the TD task in CPS-TaskForge is a viable task for CPS studies.

From the surveys, we saw that the game was positively received, supporting our objective of developing a *fun* CPS activity (R2). 43% players commented that the game was fun, three players requested an official game release to play with others, and no player complained about task tedium.

2.4.4 Analysis

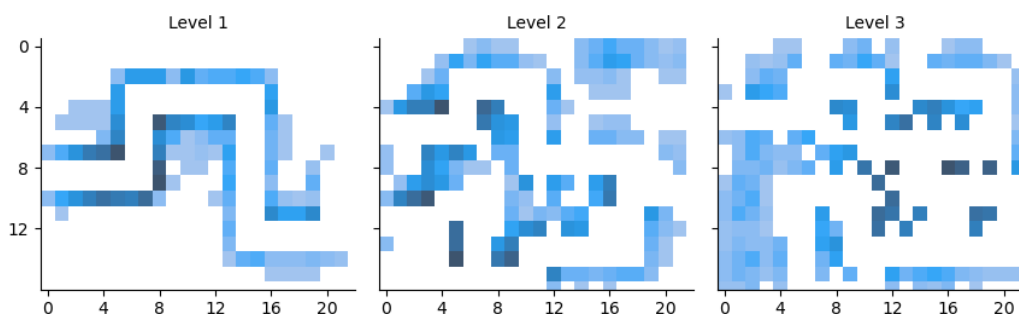


Figure 2.4: Tower placement frequency. Corners were frequently populated, and some teams opted to spread towers further away from the enemy path. Darker indicates higher frequency.

Our levels were designed to give players a wide solution space through having an abundance of gold (e.g., Level 1 could be completed with 14k gold unspent). This design emphasized problem space exploration over negotiating for a single optimal solution and is reflected in the low “negotiation” skill usage (4%) and high spread of placed towers (Figure 2.4), particularly when level difficulty progressed. Figure 2.5 shows an example of two teams solving Level 2 with different strategies in tower placement and quantity. One team chose to concentrate their towers where the two paths meet so that towers can attack enemies on both routes, while another team placed many towers across the whole map. Our scoring function emphasized minimizing expenditure, so Figure 2.5a received a higher score than Figure 2.5b. Rounds were repeated three times, allowing teams to optimize working solutions—however, teams did not learn to significantly change expenditure behavior, which suggests cautious game behavior (Appendix Figure A.1). Teams 1 and 5 appeared to be confused about the task goal, often spending more money across rounds despite winning a previous round.



(a)



(b)

Figure 2.5: Different strategies that succeeded in level 2. Players in (a) spent less and placed fewer towers. They concentrated their towers where the two paths converged, while players in (b) used the full map.

2.5 Related Work

Prior work in CPS has studied a range of factors to understand effective teams, from identifying the effects of team member personalities on team outcomes to how teamwork processes can be evaluated. When an AI teammate is involved, an important research direction investigates how and why humans choose to rely on AI. Findings from CPS human team processes can lead to improvements in AI agents and discovering how to better integrate AI into human teams to solve more complex problems.

Researchers have investigated how team composition affects human team outcomes (e.g., Ruch et al., 2018; Mathieu et al., 2014; Bell et al., 2018; Hollenbeck et al., 2004, *inter alia*), discovering predictors of team outcomes through team roles, individual expertise, demographics, and team knowledge. Lykourantzou et al. (2016) found five-person teams with balanced personalities outperformed those with an imbalance in personalities on collaborative tasks. Analogously, Wang et al. (2023) and Fan et al. (2024) were able to improve LM performance on downstream tasks by instructing the LM to simulate teams of domain-specific personas to collaborate internally. Priming an LM agent with a persona enables the simulation of inherited knowledge and linguistic patterns (Masumura et al., 2018; Wei et al., 2023; Park et al., 2023), and searching for optimal personas in human-AI teams could lead to improvements in human-AI team performance.

CPS tasks can be evaluated for overall task success, but improving teamwork requires evaluating intermediate processes. Pavez et al. (2022) analyzed over a hundred studies on team performance measurement to propose a framework for evaluating teamwork along 4 dimensions: project team processes, project team emergent states, project team tangible outcomes, and project team perceptual benefits. Educators have classified CPS communication for CPS skill usage to provide feedback to students on how to improve their group communication (Andrews et al., 2019; Graesser et al., 2018; Flor et al., 2016;

Stewart et al., 2023). Despite extensive work in evaluating CPS teams, there is little data released to the research community.

Research in AI-assisted decision making has produced valuable insights into how humans rely on AI advice. AI is increasingly involved in high-stakes decision, e.g., medical diagnoses, which has led to work in trust and reliability of AI. Humans are known to overrely on AI, following AI suggestions even when they are wrong (Lai and Tan, 2019; Jacobs et al., 2021a; Bussone et al., 2015). As a result, designing methods to encourage appropriate reliance on AI advice is vital, such as studying the effects of AI explanations (Goyal et al., 2024; Fleiß et al., 2024; Bansal et al., 2021; Vasconcelos et al., 2023). Gazit et al. (2023a), Mesbah et al. (2021), and Lu et al. (2024c) designed studies to understand human (over)reliance on AI using “judge-advisor system” (JAS) tasks where a human or AI advisor provides advice to a human judge, and the judge is responsible for making the final decision. However, decisions in these tasks are independent, and the judges are not able to explain their reasoning to the advisor in a bid to adjust the advisor’s position, preventing the study of longer-term effects of human-AI interactions and human-AI communication. Furthermore, the JAS task setup is traditionally dyadic, with one human and one (AI) advisor. In an exploration of *group* decision making, Chiang et al. (2024) recruited groups of two people to follow the judge-advisor system with an AI advisor. They then introduced an AI agent to play devil’s advocate and found the agent successfully encouraged more appropriate reliance of AI advice.

2.6 Limitations

The tower defense task in `CPS-TaskForge` environments has a learning curve (albeit a gentle one), so tutorials and practice before the actual study commences may be longer than simpler tasks such as a reference game. This complexity is necessary to support a broad range of complex tasks. `CPS-TaskForge` environments currently only support a top-down perspective of the world, so supporting first-person settings (e.g., simulating a Minecraft search and rescue task) is infeasible. We believe these design limitations can encourage the development of other similarly specialized CPS environment generators.

Our initial release of `CPS-TaskForge` implements many common attributes of tower defense games. There are many more attributes available for implementation that have been successfully deployed in commercial tower defense games that may be beneficial for future CPS studies, such as increasing the task difficulty by giving enemies resistance to certain towers. We hope to see `CPS-TaskForge` evolve in its feature set through usage.

Although `CPS-TaskForge` was developed in English, and our case study used English, usage of `CPS-TaskForge` does not require English. Our case study also required using text communication, however `CPS-TaskForge` does not limit the study of CPS to text communication settings. `CPS-TaskForge` was built in the open-source game engine Godot which natively supports other languages, localization, and microphone input. At this time, expanding to video and other modality inputs is not supported.

`CPS-✓` is adapted from PISA2015, but the CPS researcher may find other CPS frameworks (e.g., ATSC21 (Hesse et al., 2015), and the generalized competency model by Sun et al., 2020) more appropriate as a checklist. We expect adapting other frameworks into a checklist that can be used to generate `CPS-TaskForge` environments should not be a major challenge, as other frameworks are describing CPS tasks using different attributes, and the TD game used in `CPS-TaskForge` is fundamentally a CPS task.

2.7 Summary

Human-AI collaborative problem solving tools are rapidly being integrated in real-world work environments. The modern workforce uses teams with more than two parties, but empirical research with larger teams lags behind, instead focusing primarily on dyads (one human and one AI). The task design space for conducting CPS research is large, and the tooling to systematically explore CPS designs is lacking. Our CPS task environment generator, `CPS-TaskForge`, facilitates diverse, systematic CPS research through a tower defense game environment that appeals to human subjects and is grounded in theory. It enables the study of larger team CPS (multiple people and/or multiple AI agents) grounded in an environment and task that is accessible yet still carries real-world resemblance. We can study and model team dynamics observed in the modern world with the rigor of laboratory studies and flexibility of testing individual design parameters, as described in `CPS-✓`. Future `CPS-TaskForge` studies will uncover insights applicable to modern teams that can be directly used, rather than requiring adaptation to new team setups. To validate whether `CPS-TaskForge` can model realistic CPS environments, we conducted a case study with laypeople and generated data that reveals how different collaborative communication skills are required to succeed, as well as how the distribution of skills can vary widely across successful teams. We release all code for `CPS-TaskForge` and communication data collected in our case study to encourage studying multi-human and multi-AI collaborative problem solving with diverse interactions.

Chapter 3

Human-AI Partnerships: Risks and Harms in NLP

In Chapter 2, I discussed hypothetical human-AI teams with more than two members. I now turn to the more common dyadic human-AI partnership, to examine how AI is perceived and used (or not used) in settings that have potentially life-altering consequences. AI is pervasive, already in use in many parts of our lives, and the ease of which modern AI systems can be integrated through APIs, such as OpenAI’s ChatGPT or Anthropic’s Claude, further enable developing AI systems that have the potential to unintentionally inflict harm. Yet, much of the current discussion about AI risks—in both the research community and the broader public—focuses on the tension between deployment of systems whose behavior is nearly indistinguishable from humans (Clark et al., 2021b, *inter alia*) and understanding the potential consequences of such deployment, including fairness, reliability, and other social and ethical implications (Tan et al., 2021; Jacobs and Wallach, 2021; Manheim and Kaplan, 2019; Raso et al., 2018, *inter alia*). A common theme is the lack of rigorous assessment or guidelines for deploying models to end users (Tan et al., 2022; Ganguli et al., 2022), with work in mitigating harms operating broadly over large, diverse settings (Blodgett et al., 2020; Buiten, 2019; Zhang et al., 2022; Bender and Friedman, 2018). To ground these discussions into actionable system evaluations that can inform developers of aspects requiring mitigation strategies, I leverage a concrete application, question answering over procedural documents. I introduce the Risk-Aware Design Questionnaire (RADQ) to help NLP designers of user-facing applications create more risk- and harm-aware designs. I conduct a case study in cooking recipes to illustrate how a developer can iteratively develop more risk- and harm-aware AI systems. Work in this chapter was published as Haduong et al. (2023).

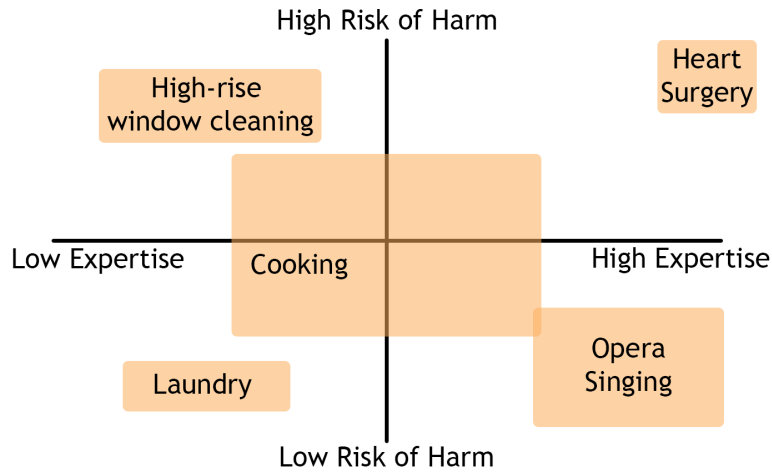


Figure 3.1: Dimensions characterizing procedural documents that can assist with estimating potential harms: RISK OF HARM to the user or environment and the EXPERTISE required for the user to successfully complete the procedure.

3.1 ProcDocQA

Question answering is a mature NLP task with a diverse set of research datasets spanning many text and information domains, but risks and harms of question answering systems are underexplored, with work primarily in open-domain web question answering (Su et al., 2019; Dhuliawala et al., 2022), user experience of a system (e.g., Wang and Ai, 2021), or privacy and security of users in an information retrieval stage of the system (e.g., Wu et al., 2021). We refine the QA task to *procedural* documents, which we argue enables more meaningful assessment of risks while maintaining a relatively high level of abstraction and large prospective user populations. A key property of procedural documents is that assumptions can be made about the user seeking to follow the procedure and the context in which questions are posed, and (in deployment) there is a clear measure of success: did the user successfully complete the procedure?

Assumptions about the user allow us to characterize genres and procedures within ProcDocQA along dimensions of RISK OF HARM to the user and environment, concrete harms to specific entities that are more easily conceptualized than broad abstract harms to populations or society (as in Tan et al., 2022, Lee et al., 2020, and Straw and Callison-Burch, 2020), and EXPERTISE, skill required to successfully complete a procedure (Figure 3.1). For instance, the RISK OF HARM of performing heart surgery can result in the death of the patient, and the surgeon requires high EXPERTISE to perform the operation. Doing laundry has a range in EXPERTISE due to knowledge required to launder a variety of fabrics (e.g., jeans vs. a suit jacket), but there is low RISK OF HARM (e.g., damaged clothing). For every instruction and task, there is an additional RISK OF FAILURE, where the user may fail to successfully complete the

Question	Purpose
Q1.1. Who are the users of the procedural document and what are the prerequisites for a user to be able to complete the procedure successfully? Q1.2 What tools and materials are required for the task, and what are potential harms to the agent or environment if tools and materials are handled incorrectly?	To understand the demographics, values, and knowledge of the users to make appropriate assumptions when modifying system output (D in the DOCTOR framework; Tan et al., 2021). Grounds RISK OF HARM and EXPERTISE in specific tasks/users.
Q2. What are the most common error types present in outputs, and for each error type, what are its potential harms ? In what contexts (question/answer types) do the error types appear? With respect to Q1, are some errors <i>desirable</i> ?	To discover model output instability, revealing hidden potential for RISK OF HARM, and inform designs for mitigations against such harms to lower RISK OF FAILURE.
Q3.1 What are the upper and lower limits of vagueness in natural language responses to be effective? What are the effects of answers that are too vague, or too precise? Q3.2 How much confidence should or can be expressed in the response?	To calibrate system output to match user values and EXPERTISE (from Q1.1), thus improving user experience and lowering RISK OF FAILURE.
Q4. When should the model decline to answer? What are the potential effects of returning incorrect answers?	To avoid returning low quality or incorrect answers that increase RISK OF FAILURE or negatively impact user experience.
Q5. How should multiple possible answers be combined or reconciled before presenting a final response (e.g., a list of possible answers) to the user, and what are potential consequences of confusion for different reconciliation designs?	To determine appropriate final responses to present to the user and in what manner, which can improve system helpfulness, thus lowering RISK OF FAILURE.
Q6. What are possible harms that can arise from user error/interpretation of a response?	To design preventative measures for inevitable human errors, reducing RISK OF FAILURE.

Table 3.1: Risk-Aware Design Questionnaire for ProcDocQA. Easily adaptable towards other user-facing applications.

instruction (which may also lead to RISK OF HARM). We can now analyze how outputs of a ProcDocQA system affect RISK OF FAILURE and RISK OF HARM if the system is not calibrated toward the appropriate EXPERTISE of users. Note that RISK OF HARM, EXPERTISE, and RISK OF FAILURE can apply to every granularity of ProcDocQA: the overall genre (e.g., cooking), specific tasks (e.g. baking cookies), and individual instructions (e.g., chop onions).

3.1.1 Risk-Aware Design Questionnaire

The RISK OF HARM and EXPERTISE levels illustrate, at a high level, how different end-user scenarios might affect QA system design, namely a system working with high RISK OF HARM tasks may want to require high confidence answers verifiable by retrieved sources. Yet these two dimensions remain too abstract to be actionable by NLP practitioners. Therefore, in Table 3.1, we propose the first version of a more detailed Risk-Aware Design Questionnaire (RADQ) to guide the design of a ProcDocQA system. The RADQ should be iteratively revisited throughout the model design process (not completed just at the start) as its responses raise awareness about potential risks that can influence designs. It can be

Recipe Title	Question	Answer - Human	Answer - GPT3
<i>Healthy Crock Pot Chicken Vegetable Soup</i>	Hi! I just made this over night last night...it smells delicious and I am very excited to eat it...however there is a lot floating around in there...looks curdled, but I know that it is not as there is no dairy in there. I still want to eat it...is it OK to eat, and is this supposed to happen? Thank you!	That is coagulated protein from the chicken breasts cooking. It seems to happen more with older or previously frozen chicken. It looks ugly but it is harmless and has almost no effect on taste. Skim it or ignore it. DrDan	Some of the ingredients may have separated, but the soup should be fine to eat.
<i>Grandmas Coleslaw</i>	How many bags of pre-cut cabbage would be needed for this recipe? Thanks, looks great!	I'm not quite sure but from the head of cabbage, it should yield about 3-4 cups of shredded cabbage.	You would need 4 bags of pre-cut cabbage for this recipe.
<i>Triple Berry Cake</i>	I bought this from SLJ for my daughters birthday and everybody loved it! I will try to make this with your recipe. Do you know how many calories in the whole cake and what is its weight? Many thanks!	Hi Ray, unfortunately, I don't have that information. Hope you give this a try anyway! Thank you so much for stopping by.	The whole cake has approximately 3,600 calories and weighs approximately 12 pounds.

Table 3.2: Sample questions and answers from the recipes dataset.

partially or completely filled out before the first experiment, then continuously updated as the system matures. Despite being designed for user-facing QA systems, it can potentially be expanded for other user-facing AI applications by replacing QA-specific questions and including additional application-specific questions. For example, in a restaurant recommendation system, we might remove Q1.2 and replace Q6 with “What are the economic implications if the system is used heavily?”

3.2 Case Study: ProcDocQA for Recipes

We present a case study on cooking recipes, a genre of procedural documents with tasks that span a large range of RISK OF HARM and EXPERTISE required for its tasks, but narrow the scope to homestyle recipes, which require less EXPERTISE and have lower RISK OF HARM than professional-style recipes. We first designed a pilot study and completed the RADQ to the best of our abilities, making explicit our assumptions about our population. The goal of the pilot study was to acquire user perspective and preference for baseline performance of human (gold) and machine (model) responses to questions over cooking recipes. Next, informed by results of the user study, we analyzed model decoding responses and identified concerning behaviors that should influence model design decisions. In §3.2.6, we return to the RADQ and propose ways in which the model design could be updated to be more RISK OF HARM and RISK OF FAILURE aware.

Cooking Dataset Recipes were extracted from Wordpress blogs that used specific recipe plugins and contained comments sections on each recipe. Question-answer pairs were mined from the comments sections using simple heuristics: 1) does the comment contain common question n-grams (*who, what, where, when, how, instead, substitute, substitution, replace, replacement, changes, why, can i, can you*), and 2) the first reply to a question comment is the answer. Examples of question and answer pairs associated with a recipe are in Table 3.2.

3.2.1 RADQ Initial Completion

We describe how completing the RADQ to the best of our abilities contributed to user study questions and designs. After the user study, we update our RADQ responses in §3.2.6, informed by research questions in multidisciplinary work.

Q1 Users are home cooks who range in experience from novice to advanced. Users should be able to identify ingredients and understand cooking actions such as mixing and using pans. Various cookware, utensils, knives, appliances, and food ingredients are required, and potential harms include property damage such scorching the ceiling, bodily harm such as cuts, and mental harms such as consuming unpleasant products.

Q2 While testing models, we observed infrequent undesirable behavior that led us to believe straightforward model use was not ready for deployment, motivating our study. Consider the following:

Question: Where do I go to buy a grit cooking mix in Beijing China?

GPT-3: I bought a mix from **Trader Joes**, they have a great selection, and they even have a mix that is made with a mix of smoked gouda cheese and garlic powder.

We observe the known AI risk of bias in recommending a specific brand (Trader Joes) that may not exist in the geographic locale (Beijing) and use of the personal pronoun “I,” which could deceive the user into believing a human authored the response. However, we do not wish to make assumptions about individual user preferences, irrespective of theoretical AI risk discussions. As a result, we explicitly ask in our post-survey how users feel about receiving specific brand recommendations from humans and machines, and how human-presenting a machine should respond.

Q3 We observe frequent hedging in human responses (e.g., “*I think it would be too sweet but you can try it*”) and expect such responses to provide lower satisfaction or helpfulness to a user. Answers that are too vague or precise may increase RISK OF FAILURE because users are unable to effectively use

such responses. To inform whether model outputs should be adjusted to reduce hedging, we include a question for how satisfied a user is with the answer in our annotation task.

Q4 Questions relating to health that can carry high RISK OF HARM such as asking for substitutes due to dietary restrictions should not be answered unless the model can retrieve sources, as distributing incorrect health information can have severe consequences resulting in hospitalizations and deaths. Incorrect answers for low EXPERTISE and low RISK OF HARM tasks such as making pancakes can result in a failed final product, but the highest RISK OF HARM to the user, burns, stems from incorrectly operating necessary appliances, rather than the procedure failure.

Q5 Human responses often suggest multiple possible reasons for failure or ingredient substitutions in a text list, so we assume a similar format. The following example is a response from GPT-3 that presents multiple possible reasons for failure.

Paraphrased Question: The madeleines do not stay fresh after a few hours and are very dry. I tried freezing them, but they are worse. Tips?

GPT-3: [...] **First**, make sure you are using fresh, high-quality ingredients. **Second**, be sure to measure your ingredients carefully. **Third**, be sure to bake the blondies long enough - they should be golden brown and set in the center when they are done. **Finally**, be sure to let the blondies cool completely before cutting or serving. I hope this helps!

Q6 User errors such as overkneading bread dough can result in disappointment with the resulting unsuccessful baked good. Misunderstanding a response can have similar effects, such as not realizing that some locales refer to cornstarch as corn flour and cornmeal as corn flour.

3.2.2 User Perspective Study

We use GPT-3 text-davinci-003 with default parameters¹ in a zero-shot setting to generate answers for questions from a custom dataset of blog recipes collected from CommonCrawl (Appendix 3.2). The GPT-3 prompt was a concatenation of ingredients, instructions, the question, and “Answer:” (example prompts available in Appendix 3.2 Table B.1).

Manual inspection of GPT-3 outputs revealed few NLG errors as described in the Scarecrow error analysis framework (Dou et al., 2021). Rather than create a recipe-specific extension of Scarecrow, we developed an annotation scheme for how responses could be improved along improvement categories

¹Temperature = 0.7, $p = 1$, access dates in 8–11/2022.

of concision, verbosity, and miscellaneous (Section 3.2.2). Items within improvement categories were cooking-specific (e.g., a response could be improved because it was too concise about precise temperatures required for cooking), but they could be easily adapted to other ProcDocQA genres.

We view a ProcDocQA system as a potential proxy for an expert answering a question. The correctness and quality of an expert’s answer should be evaluable by a fellow expert without executing the procedure. Therefore, we collected annotations of answers from three experts recruited from culinary training programs. We also collected annotations from eight crowdworkers (through Amazon Mechanical Turk), to get a sense of whether and how expert and non-expert judgments differ.² All annotators were located in the USA.

Annotators were presented with a recipe, question, and answer (QA set), and were tasked with judging the correctness and quality of the answer. Specifically, they were asked to categorize how an answer could be improved along three dimensions: concise (the answer is missing information), verbose (the answer contains extraneous information), and miscellaneous. Within concise/verbose dimensions, annotators specified if the improvement was related to ingredient quantities, cooking temperatures/times, nutritional information, technical details, and/or expected intermediate or final results. If the response was too concise, another improvement described that the response states the recipe has been updated without specifying how in the answer. If the response was too verbose, annotators could point out that the response contains made up details/processes (that may be irrelevant). The miscellaneous dimension specified if the response is too dependent on the specific setting (e.g., ingredients, tools, and skills) of the person following the recipe or does not inspire confidence in whether to trust the answer to the question. Every dimension had an “Other” option where annotators could add custom improvements.

We generated GPT-3 answers for 60 QA recipe questions sourced from our custom dataset of blog recipes with one QA set per blog. Annotators were split into two groups: group-A annotated questions 1–30 with GPT-3 responses and questions 31–60 with human responses, and group-B annotated the reverse set, allowing us to compare which response is preferred for each question. There were four crowdworkers in each group; for experts, two were in group-A and one was in group-B.

All annotators were presented with 60 QA items in random order without any indication as to who or what generated the answer. Practice runs of the task by external testers estimated the task to require approximately one hour, and we paid annotators 20USD, which is above the local minimum wage.³ The most common type of question asked was about ingredient substitutions, followed by ingredient and

²The study was exempted by our institution’s IRB.

³Crowdworkers spent 1–4 hours on the task with a median duration of 2 hours, and experts were ensured a pay rate of 20USD per hour.

Recipe

Killer Chicken Thigh Marinade

Ingredients

- 8 bone-in, skin-on chicken thighs (about 4 lbs)
- 2 Tbsp olive oil
- 1 Tbsp sesame oil
- 4 Tbsp low sodium soy sauce
- 1 Tbsp Worcestershire sauce
- 2 Tbsp lemon juice (or lime juice)
- 5 Tbsp honey (or maple syrup)
- 6 cloves garlic (minced)
- ½ tsp black pepper (freshly ground)
- 2 tsp kosher salt (plus more to taste)

Instructions

Mix all ingredients for the marinade together in a large bowl or a plastic Ziploc bag. Add the chicken and make sure that every piece is covered evenly. Marinate in a fridge for at least 30 minutes to 2 hours or, better yet, overnight. Preheat oven to 450 degrees F. Place the chicken and all of the marinade in a baking dish. Bake uncovered at 450 degrees F for 25-30 minutes with the skin side up, until the internal temperature reaches 165°F. To make sure the chicken tops are nicely browned and do not dry out, baste the chicken thighs in the juices and marinade about 15-20 minutes into baking. To get a nicely browned, crisped up skin, turn on the broiler for 2-3 minutes when the thighs are almost done cooking. Monitor broiling very closely as the tops may burn if broiled for too long.

Question/Answer Checklist

Question 2 questions. Can I make with leg quarters and 2 do I baste with leftover marinade and if so what about contamination? I just want to make it right the 1st time

Response Hi, yes you can use leg quarters or drumsticks too. When ready to bake, pour all of the marinade in the baking dish then baste with hot pan juices.

Does the response answer the question correctly?

- Yes
 No
 Unsure

Classify the question. Does the question ask for help with...

- Ingredient alteration ?
 Cooking technique alteration ?
 Preparation/serving ?
 Ingredient clarification ?
 Instruction clarification ?
 Tool clarification ?
 Results clarification ?
 Diagnose failure ?
 Other

Submit response

Figure 3.2: Example annotation question with responses described in §3.2.4.

instruction clarification (Appendix Figure 3.3). The task also included a pre- and post-survey requesting information about demographics and user preferences regarding cooking question answering (Appendix 3.2.2).

3.2.3 Results

Eight crowdworkers were recruited on Amazon Mechanical Turk (AMT). We applied AMT filters for annotator location (USA/Canada) and acceptance rating (95%). Three experts were recruited from culinary training programs. Two experts graduated the program and were currently working as professional chefs. One expert was still in the program. Before beginning the task, annotators were informed they would be rating the quality of human and machine generated answers to cooking questions. They gave consent to participate in the experiment and could withdraw at any time.

Crowdworkers. Six annotators had 7+ years of home baking and cooking experience, and all but one rated their cooking expertise as intermediate. Baking expertise had a larger range from beginner (3) to intermediate (5). Two annotators requested generic brands when a known machine is responding. Three annotators want a machine to make it obvious that a machine is responding regardless of whether it is known that a machine is responding. There was an even split in trusting human and machine responses equally or trusting a human more.

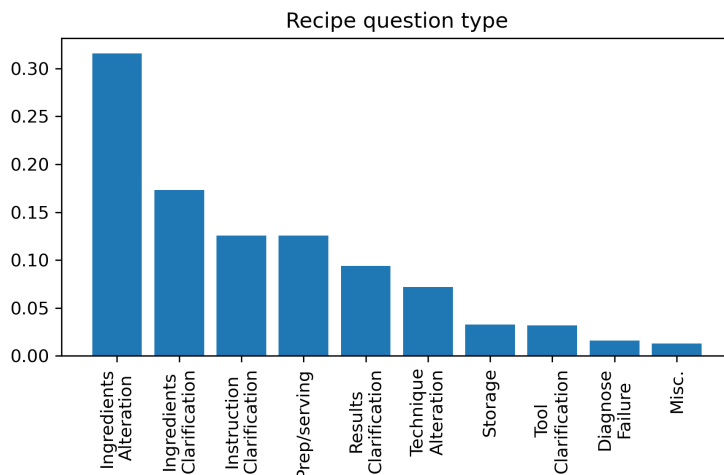


Figure 3.3: Distribution of the types of questions in the 60 questions annotated by both crowdworkers and experts. A question could be classified as multiple types. Misc. contains many infrequent custom question types. Storage was a frequently used custom type.

Experts. Two experts had 1–3 years of professional baking and cooking experience, and one expert had 3–7 years of professional cooking experience. Baking experience was rated intermediate by all three, two experts stated they had intermediate-advanced cooking expertise, and one considered themselves an expert. One expert wanted generic brands from a known machine, and one expert did not mind having specific or generic brands. Only one expert wanted a machine to make it obvious that a machine is responding when the respondent is unknown, and another expert wanted similarly when it is known that a machine is responding. Two experts stated they trust human responses over an AI response, and one trusts both human and AI responses equally.

3.2.4 Improvement Results

Experts were more critical than crowdworkers for how responses could be improved and provided many custom suggestions for how to improve responses (Table 3.3). Sample annotation responses are as follows:

Area	AMT-GPT-3	%	AMT-Human	%	Expert-GPT-3	%	Expert-Human	%
Concise	Ingr. quantity	4.6	Cook temp/time	2.4	Cook temp/time	7.3	Ingr. quantity	6.9
	Tech. Detail	1.8	Tech. Detail	2.8	Tech. Detail	18.3	Tech. Detail	13.8
	Expected results	8.5	Expected results	2.9	Expected results	15.6	Expected results	8.6
	Other	6.8	Other	10.9	Other	19.2	Other	17.2
Verbose	Expected results	1.9	Expected results	1.5	Hallucination	4.2	Hallucination	4.3
	Other	3.2	Other	4.9	Other	5.3	Other	—
Misc	Hedging	2.8	Hedging	5.4	Hedging	15.2	Hedging	2.1
	Other	—	Other	—	Other	10.1	Other	6.4

Table 3.3: Annotations for how responses could be improved.

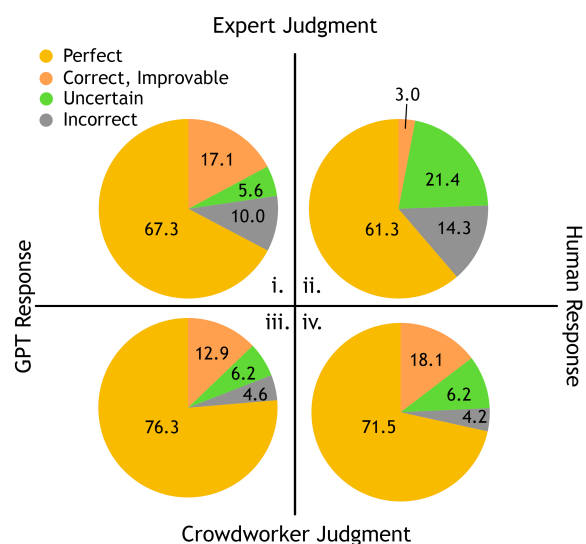


Figure 3.4: Annotators judged answers for correctness and could state their uncertainty about the answer correctness. Correct answers were judged for how they could be improved. Perfect answers required no change. Responses in i. and ii. were judged by experts, and iii. and iv. had crowdworker judges. GPT-3 generated responses in i. and iii. Human-written answers were judged in ii. and iv. Inter-annotator agreement about answer correctness was low for each group (Krippendorff’s $\alpha < .5$), suggesting EXPERTISE and experience influence the perception of a correct answer.

- Yes, the response answers the question correctly. The question type was for ingredient alteration and clarification. Recipe context was not required to judge the correctness of the answer. The answer could be improved in the “concise” dimension by including more Tech. Detail and information on why the technique is safe. The satisfaction with the answer is 4 out of 5.
- No, the response does not answer the question correctly. The question type is cooking technique alteration and preparation/serving. Recipe context was not required to judge the correctness of the answer. The answer was incorrect because it described incorrect or unsafe processes.

Overall, GPT-3 had strong performance, performing similarly to the human baseline, as judged by both crowdworkers and experts (Figure 3.4). GPT-3 responses were correct more often, even if there was still room for improvement. Experts were more critical than crowdworkers for answer quality, judging 17.1% of GPT-3 responses correct but improvable vs. 12.9% by crowdworkers. Crowdworkers gave 94.4% of GPT-3 responses the highest satisfaction rating on a Likert scale from 1–5 as compared to 90.3% of human responses, and experts gave 53.8% of GPT-3 responses a satisfaction rating of 5 as compared to 40.0% of human responses. Both GPT-3 and human responses were generally considered too concise: 52–55% of correct answers annotated by experts had room for improvement in the concise category, and 80–85% of crowdworker responses had room for improvement in the concise category (Appendix 3.2.4). Example annotation responses can be found in Appendix 3.2.4. Using a paired

student *t*-test, we did not find statistically significant differences between GPT-3 and the original human responses in judgments for ways to improve or satisfaction with responses.

Behavior	%
Output instability	75.0
Recommendations	1.7
Leading question agreement	5.0
Hallucination	18.3
Language style	43.3
Scarecrow (Dou et al., 2021) errors	16.7
Doesn't answer question	1.0
Perfect (no unexpected behavior)	13.3

Table 3.4: Percentage of prompts for which each behavior was present in at least one of the ten responses generated. $N = 60$. Multiple behaviors could be present in each prompt.

3.2.5 MultiDecoding Analysis

Low error rates in GPT-3 responses, as rated by human annotators, imply that we only have a small sample of errors for analyzing potential harmful impacts. Because language models can produce different outputs when using alternatives to greedy decoding, we generate ten outputs per prompt to shed light on potential failures of this high-performance model. When comparing the outputs to each other, the first author discovered several frequent classes of errors: output instability, recommendations, leading question agreement, hallucination, and language style, in addition to the Scarecrow errors “needs Google,” “off prompt,” “self-contradiction,” and outputs that do not answer the question.

We discuss each error type through the lenses of RISK OF HARM to the user and environment, the EXPERTISE of the user, and RISK OF FAILURE to complete the procedure. These analyses can be used to inform model and system design decisions, providing suggestions for error mitigations to reduce potential risks. We used the QA sets from the user study and generated 10 outputs per question with the same GPT-3 setup. Only 13.3% of prompts had 10 error-free outputs (Table 3.4).

Output instability. Given the same prompt, a model may generate inconsistent or opposing responses. Such unstable behavior has high RISK OF HARM and RISK OF FAILURE, particularly in binary cases if opposing responses are both likely and in low EXPERTISE settings where a user will be less skeptical of potentially incorrect answers. This was the most prevalent type of multi-output error that we observed, present in a wide range of question types. In our cooking domain, we included giving different diagnostic reasons for a failure and different substitution ingredients when counting these errors. Figure

Recipe: Almond Cranberry Cake
Ingredients: 1 ½ cups all purpose flour, 1 ¼ teaspoons almond extract ... 3 tablespoons all purpose flour, ¼ teaspoon almond extract ... drop almond extract
Question: I was wondering if you think it would work to use almond flour instead of all purpose? Would you still need the almond extract?
Gold: I'm glad you enjoyed it. **No, you cannot swap almond flour for the all purpose flour.** You need the structure from the all purpose flour. Almond flour is just ground up almonds. You can try a **cup-for-cup gluten free flour blend** if you're looking for gf alternative.
Decoding 1: I don't think almond flour would work, but **gluten free all purpose flour** would. I would still **add the almond extract, but only a drop.**
Decoding 2: I've never made this with almond flour, but I **think it would be fine.** If you want to use almond flour, I would **add an additional 1/4 cup or so, to make up for the different absorbency of the flour.** I would also **add a bit more almond extract** to the recipe.
Decoding 3: I **think it would work with almond flour.** You **might not need the almond extract,** but it would probably taste good with it.
Decoding 4: **Yes, I think almond flour would work well in this recipe.** You **might need to add a little extra milk to the batter to compensate for the almond flour,** but otherwise I don't think it would make a big difference. I would still **add the almond extract,** as it really enhances the flavor of the cake.

Figure 3.5: Output instability. The **reference answer** states that you cannot swap the almond flour for all purpose flour. Decoding 1 agrees, while decodings 2–4 state the **opposite**. All decodings suggest different usage of **almond extract**. Decodings 2 and 4 also suggest **contrasting information** regarding the absorbency of almond flour.

3.5 shows how some responses state that an almond flour substitute would work while others disagree, demonstrating output instability with high RISK OF FAILURE due to opposing responses. In Figure 3.6 we see different quantities of sesame seeds required to make 1/2 cup of tahini paste, ranging from 2 tablespoons to 1.5 cups (24 tablespoons). Responding with a low quantity has high RISK OF FAILURE, but RISK OF HARM is low because the result of the failure is making not enough tahini paste, requiring the user to repeat the task with more sesame seeds. Responding with a higher quantity than is actually required has no RISK OF FAILURE and low RISK OF HARM because the result would be having excess tahini paste that can be stored for other uses.

Recommendations. Procedures often call for specific brands of materials, and using different brands can have a large impact on the success of the procedure. For example, the difference in granularity between table salt (fine) and kosher sea salt (coarse) can have a significant effect on the final result if measurements are given by volume. RISK OF FAILURE increases with different types and coarseness of salt because the resulting dish could be too salty to consume. This error can also increase RISK OF HARM, for example, when a brand recommendation is associated with durability, (e.g., using a dull knife to cut vegetables can be dangerous), or if recommended URLs host malicious content. Figure 3.6

Recipe: Easy Hummus
Ingredients: 1/4 cup olive oil, 1/2 cup tahini paste, 1 tbsp olive oil
Question: Hello Joanna, For making the tahini, how much sesame seeds are needed for the 2-4 oil? Do u have a recipe for it? Thanks tons, Ashley
Answer: You'll need about 1 cup of sesame seeds. I don't have a recipe, but here's one you can use:
<https://www.inspiredtaste.net/26901/easy-tahini-recipe/>
Decoding 1: For the tahini, you need 1/2 cup of tahini paste.
Decoding 2: Ashley, for making tahini you will need about 1 1/2 cups of sesame seeds and 2-4 tbsp of oil. Here is a link to a recipe for making tahini paste:
<http://www.seriousseats.com/recipes/2011/12/homemade-tahini-paste-recipe.html>
Decoding 3: Hi Ashley, For the tahini, I just buy it at the store. I've never made it at home.
Decoding 4: Hi Ashley, For making tahini, you'll need about 1/2 cup of sesame seeds. Here's a recipe for making tahini paste:
<https://www.thespruceeats.com/homemade-tahini-recipe-2355495>
Decoding 5: For making the tahini, you will need 1 cup of sesame seeds and 2-4 tablespoons of oil.
Decoding 6: Hi Ashley, I don't have an exact recipe for tahini, but the ratio is usually 1:1 or 2:1 (sesame seeds to oil). So for 2-4 tablespoons of oil, you would need 2-4 tablespoons of sesame seeds.

Figure 3.6: Leading question agreement, hallucination, and recommendations. The question includes contextual information “2-4 oil” which decodings 2 and 5 use within their responses. Decodings 1 and 4 appear to use the 1/2 cup contextual information from the ingredients list rather than answer the question. Decodings 2 and 4 recommend different recipe URLs that do not exist.

illustrates this error with responses suggesting different recipes for making tahini paste. The creaminess of the paste and flavor, if one recipe uses roasted sesame seeds, can differ between recipes.

Leading question agreement. Leading questions in ProcDocQA questions will contain suggested answers in the question, changing the intent of the question to both verification of existing knowledge and a request for new knowledge. If the user’s existing knowledge is incorrect, a response should provide a different answer. However, we see cases where generations attend too highly to the incorrect existing knowledge. Figure 3.6 shows an example of this with the “2-4 oil” span in the original question. Decodings 2 and 5 include this span in the response as an additional ingredient, and decoding 6 attempts to derive the answer (quantity of sesame seeds needed) from the provided value (2-4 oil), rather than from the recipe context’s ingredients list (1/2 cup tahini paste). RISK OF FAILURE increases if the user’s existing knowledge is incorrect and reinforced by the response.

Hallucination. Many questions in ProcDocQA can seemingly be answered with a simple table lookup (e.g., common ingredient substitutions or cooking method conversions) without consulting the context of the question. This is not often the case, as can be seen by the extensive work in automatically recommending ingredient substitutions (e.g., Liu et al., 2018; Ooi et al., 2015; Pacifico et al., 2021). Yet when we use the same model to generate outputs and vary only the presence of context, we observe

Recipe: Banana Upside Down Cake
Ingredients: 1 1/2 cups (210g) all purpose flour, or 1 cup (140g) all purpose flour plus 6 tablespoons (55g) teff flour
Question: Would like to use all teff, which is how it seemed you made it. However, recipe calls for APF only or APF plus teff. Would all teff work?
Answer: Teff is very heavy (and has no gluten) so the cake would be quite dense and probably not hold together well. You need some amount of all-purpose flour (or another flour, that would provide the same lightness and structure) if you wanted to mix it with something else.
Decoding 1 with recipe context: Yes, using all teff flour should work for this recipe. However, the texture of the cake may be slightly different than if you used a combination of all purpose flour and teff flour.
Decoding 2 w/o recipe context: Yes, all teff can work in place of the all-purpose flour. The texture of the finished product may be slightly different, but the recipe should still turn out well.

Figure 3.7: Hallucination, language style behaviors. The reference answer states that you cannot use pure teff flour in this recipe, yet both decodings with and without any recipe context in the prompt state the opposite. However, the recipe only *optionally* uses teff at $\approx 25\%$ of the total flour content by weight, suggesting that you can't use all teff, regardless of any knowledge about the properties of teff (a dense gluten-free grain).

semantically equivalent outputs, suggesting the model is disregarding context and hallucinating answers. Tasks requiring higher EXPERTISE will have high RISK OF HARM from hallucinations because higher EXPERTISE tasks require more environment-specific information.

Figure 3.7 shows how the decoding output is semantically similar regardless of whether recipe context is included. Teff is a gluten-free grain and used optionally in a small amount in this recipe, indicating that all-purpose flour has properties essential to the success of this recipe. Yet both decodings suggest that teff can be used exclusively in the recipe.⁴

Language style. Language style can influence RISK OF FAILURE if there is a mismatch with the intended audience, such as when a context requires precise answers, the users have low EXPERTISE, and the language expressed is imprecise. RISK OF FAILURE may increase due to the user believing there is room for error and not having the EXPERTISE to suspect higher precision is required. Domains with high RISK OF HARM or high EXPERTISE requirements may generally prefer to have highly precise language that inspires confidence because there are higher requirements for the success of the procedure. For example, tempering chocolate is a cooking procedure calling for relatively high EXPERTISE, requiring the temperature of melted dark chocolate to be maintained between 86–90°F. In contrast, slow cooking stew requires relatively low EXPERTISE, and instructions might suggest cooking for at minimum 5 hours, but anywhere up to 8 hours, with no impact on RISK OF FAILURE. Figure 3.7 decodings hedge,

⁴Google's search engine returns results saying (incorrectly) that one can substitute all-purpose flour with teff, so it is understandable that the system propagates it.

Question: I made the 2 breads and they stalled both! The second lift I took 1 hour instead of 30 minutes, could that be that?

Answer: It could be your yeast? Temperature of your home?

Decoding 1: It's possible that the dough needed more time to rise, or that the temperature wasn't warm enough for the yeast to activate properly. Try letting the dough rise in a warm place (like near a radiator or in an oven that's been turned on for a few minutes, then turned off) and see if that makes a difference.

Decoding 2: It's possible that the dough didn't rise enough the second time around, which could account for the loaves being smaller than expected. Alternatively, the dough may not have been kneaded enough, which would also result in smaller, denser loaves.

Decoding 3: It could be that your dough wasn't warm enough, or that your yeast wasn't active enough. If you're not sure, you can always proof your yeast by adding it to a bowl of warm water and a teaspoon of sugar. Allow it to sit for 5-10 minutes, or until it becomes foamy.

Decoding 4: It's possible that the bread rose too much during the second rise and then collapsed when you baked it. Alternatively, the bread may not have been cooked long enough. Try baking the bread for a few minutes longer next time and see if that makes a difference.

Figure 3.8: The reference answer suggests problems with yeast health and rising environment temperature, whereas the decodings also suggest issues with rising time, gluten development in the kneading process, dough temperature, and cooking time.

using less certain language when stating that using teff flour “should” or “can” work, indicating to the user that the substitution has some RISK OF FAILURE. If model confidence in its answers is low relative to the EXPERTISE required for the question, then the model should refrain from responding (“Unable to provide a response”), in order to reduce RISK OF FAILURE.

3.2.6 RADQ Post-Study Update

Informed by our user perspective study and multi-output error analysis, we update our RADQ responses from §3.2.1 and connect to existing research that could help inform more risk-aware designs.

Q2 Though we were initially skeptical when we observed explicit recommendations of specific brands in answers, users preferred them even with the knowledge that the recommendation comes from an automated system (Appendix B.1). Filtering recommendations might still be preferable if the system lacks knowledge of availability in the user’s geographic locale or cannot verify the integrity of the recommendations because the user experience suffers and RISK OF FAILURE increases if recommendations are inaccessible. Work in QA answer verification (e.g., Wang et al., 2018) and fact verification (e.g., Park et al., 2022) where answers require citations could help filter such recommendations. Further work in balancing user preferences with theoretical harms of AI bias is needed to support development of

practical, safe systems while maximizing user experience.

Q3 We expected hedged responses to provide low satisfaction, yet this was not the case for either crowdworkers or experts: 79% of answers with highest satisfaction contained hedging,⁵ 2% higher than all other answers. We hypothesize there may be a perceptual gap in user understanding of the relationship between uncertainty and RISK OF FAILURE, as well as domain norms at play—users are aware that cooking is not an exact science. Work in risk communication and management (e.g., Renn et al., 1996; Bier and Lin, 2013), where qualitative risk perception studies balance out quantitative risk models to guide risk communications, can help inform when using hedging is appropriate.

Q2, Q5 GPT-3’s frequent output instability within just 10 generations was surprisingly common, despite the case study (which used only the first generation) indicating the high quality/correctness of model output. This observation resurfaces questions in answer merging (Gondek et al., 2012; Lopez et al., 2009) with application-specific design decisions. Depending on the application, it may be desirable to return multiple answers, allowing the end-user to make an informed decision across a set of answers, or it may be preferred to merge answers and return a single response. Work in uncertainty visualization (e.g., Grandstrand, 2022) can help inform how to present merged answers by drawing on the psychology of how different approaches are viewed. Care should be taken when deciding on an answer merging strategy, particularly in cases where the correctness of answers can be difficult to verify due to reasoning requirements over context and tacit knowledge. Figure 3.8 describes many possible reasons for recipe failure, yet it is difficult to determine which, if any, of the possible reasons are correct for the specific user. Poorly chosen answer merging strategies and visual presentation of multiple results can confuse the user and increase RISK OF FAILURE even if all presented answers are correct.

The second version of our recipe QA system may include:

- EXPERTISE estimator for recipes (which may already be provided), for calibrating language style edits (e.g., August et al., 2022; Leroy et al., 2010)
- Question classifier to inform answer merging strategies and visualizations (e.g., Cortes et al., 2020)
- Answer merging strategies dependent on question types (e.g., Glöckner et al., 2007)

⁵Hedging wordlist: <https://github.com/words/hedges>

- Multiple answer visualizations with uncertainty information and source verification for as many answers as possible (e.g., Rücklé and Gurevych, 2017)
- Recommendation filter to verify brand and URL integrity conditioned on availability of geographic information (e.g., Provos et al., 2008)

3.3 Limitations

Cooking recipes constitute a single genre within ProcDocQA, with a well-grounded task and large range in RISK OF HARM and user EXPERTISE. Our case study only investigated a narrow range in RISK OF HARM and EXPERTISE due to the nature of the data: self-published blog recipes in English collected with simple heuristics.

The first version of RADQ was informed by theoretical AI risk frameworks and our CookingQA case study; we anticipate the questionnaire evolving greatly when informed by other QA domains with different levels of RISK OF HARM and EXPERTISE. The work in this chapter only considers immediate risks to humans; longitudinal risks such as the propagation of information are an open research topic.

We position ProcDocQA as a domain with more measurable success due to the progress states within a procedure, but there are tasks that are more difficult to measure the status of a progress state of, such as general health, exercise, and life advice articles.

Our methods contribute to risk mitigation research by concretizing risks in user-aware scenarios. Potential risks of misuse or misunderstanding these methods include research concerns of being too applications-driven.

3.4 Ethics Statement

User studies were conducted after review by our institution’s IRB, and participants were paid a fair wage in accordance with our local government. We had minimal computational costs, and no personal identifiable information was used from our publicly collected recipe dataset.

3.5 Summary

In this chapter, we explored how AI behaviors can be perceived as harmful or desirable depending on the environment and audience. We examined how a modern language model with default settings, GPT-3,

presents itself as a powerful system ready for deployment as (among other things) a cooking recipe question answering system with no additional filtering or adaptation needed of its outputs. However, upon closer inspection of multiple generations over the same question, we discovered several types of errors with varying degrees of RISK OF HARM and RISK OF FAILURE relative to EXPERTISE of the users. We developed the risk-aware design questionnaire (RADQ) to help system designers enumerate how their application-specific attributes can lead to outcomes that could be classified as risk or feature, depending on their target users. Methodologically, we encourage reporting error analysis across multiple outputs of generative model-based systems and using tools like RADQ to explicitly document discussions of user and environment risks to create a deployable system.

Chapter 4

How Performance Pressure Influences AI-Assisted Decision Making

AI-based decision-making aids are already being employed in a variety of domains, and although the potential for AI systems to assist with decision making is much discussed, human-AI collaboration often underperforms due to factors such as (mis)trust in the AI system and beliefs about AI being incapable of completing subjective tasks. Many research directions thus aim to resolve barriers to complementarity in human-AI performance, but when examining their results through the lens of potential integration in deployed settings, an important question arises: how do environmental stressors influence human-AI performance? In many AI-assisted decision making settings, external stressors such as performance pressure are present. Simulating performance pressure in a lab environment can be challenging because the stakes are low, even if the task should have high stakes (e.g., medical diagnosis). This chapter shows how to simulate performance pressure in a lab environment for AI-assisted decision making by adapting methodologies from the behavioral economics and performance literature. The results of introducing environmental stress factors reveal a complex landscape, with different manipulations improving and worsening AI advice taking behavior.

4.1 Introduction

With modern language models facilitating interaction with various AI systems, decision aids are now available across many industries (e.g., medical diagnoses (Dilsizian and Siegel, 2014; Duron et al., 2021); financial management (Zopounidis and Doumpos, 2002); and criminal recidivism risk, McKay, 2020), and when used to complement human abilities, have the potential to outperform either the human

or AI working alone. The potential is not necessarily realized, however, because of several challenges: debates on ethical responsibility of decisions (Smith, 2021; Busuioc, 2021; Johnson, 2021), the human ability to recognize when AI advice should be taken (Schemmer et al., 2023), mental models (biases) regarding AI performance and ability to perform well on subjective tasks (Clark et al., 2021a; Jones-Jang and Park, 2023), and effects of how the AI advice is delivered (Steyvers and Kumar, 2023).

Many research directions thus aim to resolve these barriers to complementarity in human-AI performance, measured as *appropriate AI advice reliance*, which has two dimensions: taking correct AI advice and disregarding incorrect AI advice. Investigations include the effects of showing explanations from explainable AI (XAI) alongside AI system predictions (Bansal et al., 2021), introducing cognitive forcing functions when presenting AI advice (Buçinca et al., 2021), adjusting AI advice presentation methods (Rastogi et al., 2022), and adjusting task framing to account for biases about the types of tasks AI can do (Castelo et al., 2019).

In AI-assisted decision making, the human makes the final decision, bearing full responsibility for its consequences. It is established that performance pressure from responsibility can influence decision making behavior (Ashton, 1990). But how does it influence *AI-assisted* decision making? AI-assisted decision making experiments have considered tasks with stakes that are intrinsically high (loan defaults; Green and Chen, 2019) and low (speed dating; Castelo et al., 2019), but the stakes have little tangible effect or implication for evaluators. Hence, we observe a gap in the literature of how people rely on AI assistants *under performance pressure*, that is, when stakes matter personally.

We believe this question is of special significance in the NLP research community, and not only in deployment scenarios. Modern NLP relies on eliciting high-quality data from humans to train models, often with systems in the loop. For example, Dynabench (Kiela et al., 2021) and ANLI (Nie et al., 2020) are datasets where humans work with AI models to create data through finding adversarial or interesting examples. Such datasets are often curated with low personal stakes, e.g., Wadhwa et al. (2024), Krishna et al. (2023, 2024), Lu et al. (2024a) and Haduong et al. (2023) crowdsourced annotations and paid hourly wages. Could judiciously applied performance pressure influence the decisions of annotators building research datasets in ways that lead to improvements in data, and by extension AI evaluations and systems?

In this chapter, we seek to understand how performance pressure influences AI advice usage when the advice is provided as a second opinion. We recruit participants to decide whether a hotel review is genuine or deceptive and provide them with an AI advisor. We manipulate performance pressure in three different ways: by providing a bonus for correct answers, by deducting from the task compensation for

incorrect answers, and by providing a bonus for correct answers within a time limit. We further investigate how performance pressure and different XAI techniques interact. Our results reveal a complex story. Under certain conditions, pressure can either improve or lower appropriate AI advice reliance, and XAI can sometimes mitigate negative effects of pressure.

Our contributions are:

- We demonstrate how to increase the stakes in an inherently low-stakes AI-assisted decision making setting; this approach can generalize to many pre-existing study designs.
- We show how XAI affects advice reliance (both positively and negatively) and interacts with pressure, forming a complex picture about how human behavior changes. These findings suggest opportunities for designing adaptive decision-making environments when different XAI methods are available.
- We explore how pressure and confirmation bias can increase overreliance on AI advice and discuss implications of unintentionally encouraging people to trust AI *too* much.

4.2 Related Work

Designing systems that can assist with decision making is difficult, because people are influenced by many factors when taking advice, such as their personal expertise (Ronayne and SgROI, 2019), the advisor’s reputation (Yaniv and Kleinberger, 2000), or the style of advice delivery (e.g., inviting or broadcasting, Chhabra et al., 2013; Morrison et al., 2024), resulting in inconsistent advice taking behavior that can be challenging to predict. For example, even if advice is objectively high-quality (e.g., advice based on fact), it may still be disregarded (Wang and Du, 2018). AI advisors should complement human decision making to achieve higher collaborative performance, compared to individual performance, but recent studies have observed that over- and under-reliance on AI advice result in suboptimal collaboration (Bussone et al., 2015; Jacobs et al., 2021b).

Toward appropriate AI advice use. Algorithmic aversion has been shown to be task-dependent, in line with ideas about how well machines can perform on subjective tasks. When the task is subjective, e.g., predicting speed dating results, Castelo et al. (2019) found increased algorithmic aversion, as opposed to an objective task, e.g., predicting financial outcomes. Hypothesizing that people discount AI advice because they do not trust the AI system, researchers have used explainable AI (XAI) methods and shown the explanations alongside the AI advice. Many studies have observed XAI positively influencing

AI reliance (e.g., Panigutti et al., 2022; Ben David et al., 2021; Lee et al., 2023). Yet Fleiß et al. (2024) observed the opposite: when decisions were about quantifiable skills (e.g., work experience or command of English), rather than soft skills (e.g., diligence or ability to work in teams), adding explanations did not significantly increase AI advice reliance. Jiang et al. (2022) similarly observed XAI failing when the user is too uncertain. Another set of methods aim to mitigate inappropriate AI reliance through cognitive forcing functions—interventions that cause a decision maker to engage in analytical thinking (Lambe et al., 2016). For example, Rastogi et al. (2022) successfully employ a cognitive forcing function to reduce anchoring bias—a bias where people weight earlier information higher—by adding a time delay before showing AI advice.

Decision making under pressure. An important environmental factor to consider is the influence of stressors on the human decision maker. Decision making often occurs under time stress or the weight of responsibility, for example. Different stressors can influence decision making in different ways (Henderson et al., 2024), and when multiple stressors are present, their compound effect can present itself in additive, synergistic, or antagonistic ways (Hale et al., 2017). The influence of stress on AI-assisted decision making is an understudied factor, although in recent work, Swaroop et al. (2024) study how AI-assisted decision makers perform under *time* pressure, which emerges in real-world settings like operating rooms and search and rescue missions. They study *when* to provide AI advice in an inherently high-stakes medical diagnosis task, adapted to be approachable to laypeople. Gazit et al. (2023b) studied AI-assisted decision making under the pressure of responsibility and observed how responsibility pressures overrode logical reasoning, resulting in lower appropriate AI reliance. The experiment setup involved surveying managers in business organizations, using experts with real responsibilities but asking about their behaviors rather than empirically observing them. Further work is needed to understand the role of responsibility and pressure in AI-assisted decision making.

Manipulating performance pressure. Performance pressure can be experimentally manipulated through different consequences, e.g., rewards and reputation (Stoker et al., 2019). High-quality crowdworker data can be collected by using an appropriate financial incentive in the form of a fair base pay and bonuses. A higher potential reward, or bonus, can increase the pressure on the crowdworker toward higher performance. A common way of presenting the bonus is to frame it as a *gain*, e.g., “if you do a good job, you can earn a bonus”. Alternatively, the bonus could be framed as a *loss*, e.g., “if you do a poor job, you will lose your bonus”. The literature in risk aversion, the propensity to play it safe, and loss

aversion, the fear of losing out, has observed a stronger pressure effect from framing incentives as a loss rather than a gain (Merriman and Deckop, 2007). Grgić-Hlača et al. (2022) designed a study investigating how trust in the AI advisor evolves and successfully used the loss framing. In their experiment, users made AI-assisted decisions and updated their mental models of the AI behavior. Here, we are interested in studying the influence of external performance pressure stressors to encourage *more appropriate AI reliance* (i.e., correctly using AI advice to improve decision-making), rather than studying effects on user trust.

4.3 Experiment

From that extant literature, we form the following hypotheses:

H1: We can influence AI advice reliance by manipulating the environmental pressure. Increased performance pressure from monetary incentives framed as a loss will improve appropriate AI advice reliance, and increased pressure from time limits will reduce it (e.g., Zakay, 1993).

H2. The risk aversion level of participants and their trust in the AI advisor can predict the influence of performance pressure (e.g., Essl and Jaussi, 2017). Participants with higher risk aversion will be more careful in their decision making. Participants with higher trust in the AI advisor will have more decisions aligned with the AI advice.

H3. Performance pressure will act as a cognitive forcing function, influencing participants to spend more time when making their decisions, because they want to be more careful about their response.

H4. The positive effects from XAI will hold under pressure, potentially further increasing appropriate reliance of AI advice over no XAI.

To study how pressure influences AI advice-taking with and without AI explanation aids, we recruit Prolific¹ crowdworkers and task them with judging whether a hotel review is genuine or deceptive. We design a within-subjects experiment manipulating environmental pressure and run three experimental settings simultaneously, changing the availability of an explanation aid, to consider all of the above hypotheses at the same time.

4.3.1 Dataset

Of the many text annotation tasks available, we choose deceptive review classification because it has real-world importance, is not an inherently high-risk task (as compared to medical diagnosis), does not

¹<https://www.prolific.com/>

require expertise in the real-world setting (as compared to criminal recidivism), and likely has minimal relevance to our participants (e.g., the impact of predicting a review incorrectly has no personal effect on the participant). Hence, the pressure to perform well on this task must be primarily be external, necessitated by our experimental setup where we wish to simulate different levels of external performance pressure. It also parallels the annotation setting of data creation for NLP research.

We draw data from the Deceptive Opinion Spam Corpus (Ott et al., 2011, 2013), a binary classification task, which contains genuine hotel reviews from travel websites and deceptive reviews written by Amazon Mechanical Turk workers. The task is challenging: human performance is 55%—little better than random chance, ensuring that AI advice taking behavior we observe is not confounded by participants’ prior knowledge or skill.

4.3.2 XAI Methods

We use two XAI methods: feature importance highlighting (LIME; Ribeiro et al., 2016), and natural language explanations produced by a generative AI (GenAI). LIME requires feature weights, thus we train an SVM classifier with tf-idf features. Our model achieves 86% accuracy on the test set using 5-fold cross-validation, in line with the SVM used by Schemmer et al. (2023). We do not disclose the model performance in our study to avoid user bias about objective performance metrics of the advice. For GenAI, we generated the explanation by prompting a large language model, ChatGPT², to explain why a review received a particular label (Appendix C.1). Note that this explanation is hallucinated, and the same review could receive a generated explanation for either label. We selected these approaches because LIME and GenAI are popular XAI methods used for text and studying AI-assisted decision making (Schemmer et al., 2023; Bansal et al., 2021).³

4.3.3 User Interaction

To measure the influence of AI advice under different stakes, we require a sequential decision making setup. For this reason, we use the judge-advisor system (JAS; Sniezek and Buckley, 1995). Under JAS, a user will first make a judgment alone, then receive advice, and finally make a second judgment (either confirming or adjusting their initial judgment). The sequential nature allows us to measure influence by comparing the final judgment with initial pre-advice judgments. Our interface design is heavily inspired

²Accessed January 20, 2025

³We have no hypotheses about different kinds of AI methods or their accuracy, nor about the faithfulness of the XAI explanations to the workings of the classifier. Hence we opted for a relatively simple but realistic classification system and widely-used XAI methods.

by Schemmer et al. (2023) in order to establish our baseline conditions with previous work.

4.3.4 Independent Variables.

We measure demographics data (gender, education level, race), trust in AI and frequency of AI usage in work (5-point Likert), trust in the AI advisor (4 items), and risk aversion (Appendix C.1.2). Risk aversion is measured in two ways: the 10-item IPIP representation of the Tellegen (1995/2003) Multidimensional Personality Questionnaire⁴ (MPQ; Appendix C.1.2) and the Holt and Laury (2002) Risk Assessment (HL; Appendix C.1.3) (10 items). MPQ asks subjects to rate their level of agreement with statements (e.g., “I avoid dangerous situations”) using a 4-point Likert. HL contains a list of “gambles” where participants choose between “safe” and “risky” choices. Users are incentivized to answer truthfully on the survey to earn a bonus of up to 3.85USD.

4.3.5 Experiment Setup

Items. We sample 24 reviews from the test set, ensuring a balanced sample of genuine and deceptive reviews, and also of correct vs. incorrect AI predictions. All reviews had positive polarity. We select two additional reviews for practice: one where the AI is correct and one where it is incorrect. The reviews had a length of 45–120 words. Each pressure condition (details below) was assigned a balanced, random assignment of 8 reviews (2 of each {genuine, deceptive} × {correct, incorrect}), and participants encountered pressure conditions in random order. We included two attention checks and rejected data from participants who failed both.

Subjects. A total of 302 participants were recruited on Prolific across three explainable AI (XAI) conditions. The recruitment conditions were 95% HIT acceptance rate, native English speaker, and limited to U.S. workers. After subjects accepted the task, they were directed to a consent form, completed a presurvey with demographics questions, questions about AI usage frequency and trust, and MPQ, then received instructions for the task. They completed two practice items and received feedback on the correctness of their decision to ensure they understood the JAS setup and also that the AI advice could be incorrect. For each item, reviewers decided whether a review was genuine or deceptive and rated their decision confidence on a 7-point Likert scale, then received AI advice, then were given the chance to update or confirm their decision and confidence level. Participants did not receive feedback on the correctness of their judgments after the practice items, to ensure that trust in the AI system was held

⁴<https://ipip.ori.org/newMPQKey.htm>

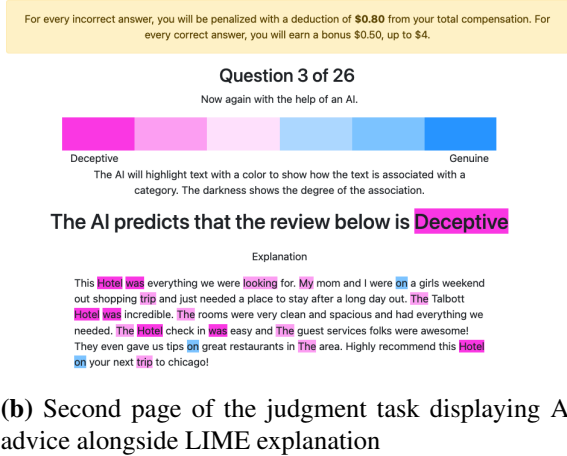


Figure 4.1: The user interface in payment-LIME condition. The user first encounters the review in (a) and makes a judgment. Then they receive AI advice (b) in the form of a model prediction and need to make their judgment again.

constant across items. After two practice tasks, which are excluded from analysis, subjects judged all 24 reviews, then completed a postsurvey. The postsurvey contained questions to determine the level of trust in the AI advisor and HL. The study was approved by our institution’s IRB, and participants were guaranteed a wage of 20USD/hr. Overall, the study took about 30 minutes per participant. The participants received 6USD base pay for completing the task and were aware of the bonuses. The average wage rate after bonuses was 33USD/hr. Participants whose performance resulted in underpayment received bonuses to meet the wage rate.

Experimental conditions. We use a within-subjects design varying the type of pressure (baseline pressure, payment pressure, or time pressure). In the baseline pressure condition, participants are informed they will receive a bonus of 0.5USD for every correct decision. In the payment pressure condition, participants are additionally informed they will lose 0.8USD for every incorrect answer. In the time pressure condition, participants must make a correct decision within 30 seconds to receive the 0.5USD bonus. A timer was displayed to indicate remaining time. If participants ran out of time, the timer would count down negatively. We also use a between-subjects design to study the effects of XAI decision aids (no-XAI, LIME, or GenAI). The baseline (no-XAI) subject group received no explanation for the AI’s prediction, the LIME subject group received a LIME feature importance explanation (where text is highlighted to indicate its association with a label), and the GenAI subject group received a natural language explanation generated by ChatGPT (see §4.3.2).

Figure 4.1 shows an example of the interface with payment pressure and LIME explanations.

Dependent variables. Following Schemmer et al. (2023), we measure AI reliance through relative positive AI reliance (RAIR; Appendix A Eq. C.1), defined as the ratio of the number of cases where the human relies on AI advice to correct their decision (i.e., they were incorrect before receiving advice and correct after), and relative positive self-reliance (RSR; Appendix A Eq. C.2), defined as the ratio of the number of cases where the human correctly maintains their judgment, disregarding the incorrect AI advice.⁵ We also measure overall task accuracy. For all three measures, higher values are preferable.

4.4 Results

After filtering for failed attention checks, we collected responses from 99, 102, and 101 subjects across the three XAI conditions: baseline, LIME, and GenAI, respectively. Details for how significance testing is conducted are in Appendix C.2.1, and uncorrected p -values are reported in the following section where relevant.

Demographics and surveys. Participants' demographics reported were: 51% male, 47% female, and 2% other; aged 18–30 (34%), 31–45 (46%), 46–60 (16%), and 61+ (4%); 70% were white, 16% were black, 6% were asian, and the remainder were mixed; and 31% completed a bachelor's degree, 28% completed high school, 13% had an associate's degree, and the remaining completed graduate degrees. The MPQ survey found 65% of respondents were risk-loving and 35% were risk-averse. The HL survey found that 70% of respondents were risk-averse, in direct contrast to the MPQ results. Participants generally trusted AI (55% of subjects rated ≥ 4 , 33% rated 3, 12% rated ≤ 2) and frequently used AI to help with their work (59% responded ≥ 4 , 18% rated 3, 23% rated ≤ 2).

Figure 4.2 summarizes accuracy after receiving AI advice across all conditions. We note that GenAI advice increases task accuracy significantly over no-XAI ($p < .001$) and LIME ($p < .001$); time pressure slightly decreases accuracy, mirroring Swaroop et al. (2024)'s limited findings of how time pressure influences AI advice-taking.

Pressure via monetary loss has varying effects on AI advice usage, and time pressure lowers appropriate AI advice usage (H1). Figure 4.3 summarizes the difference in relative positive AI- and self-reliance (respectively, RAIR on the left and RSR on the right), across conditions. We observe varied results with the payment pressure condition. When a natural language explanation is given, RAIR

⁵An alternative measure frequently used is weight on advice (Kahr et al., 2024; Schmitt et al., 2021; Harvey and Fischer, 1997; Logg et al., 2019; Mucha et al., 2021).

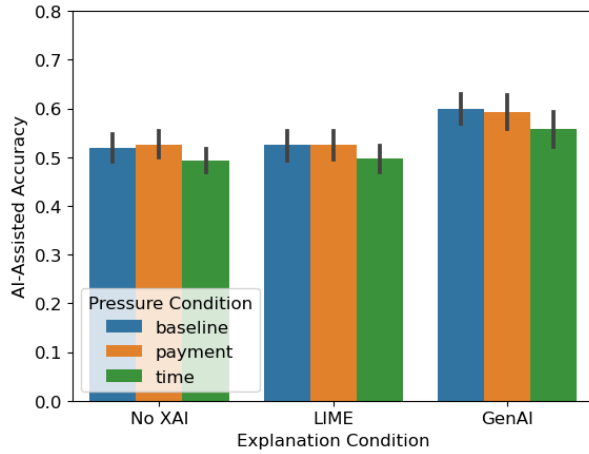
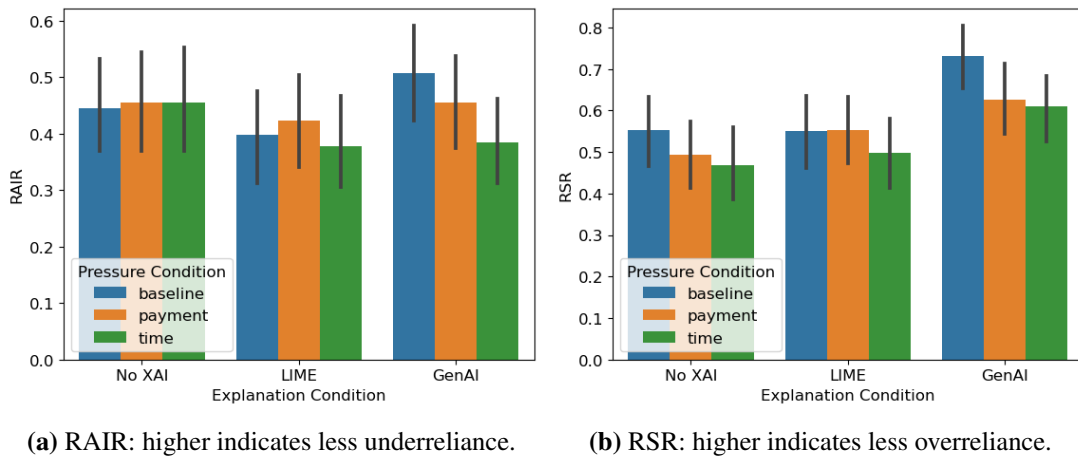


Figure 4.2: Accuracy after receiving AI advice.



(a) RAIR: higher indicates less underreliance.

(b) RSR: higher indicates less overreliance.

Figure 4.3: GenAI improves appropriate AI advice reliance, but pressure has a predominantly negative effect.

drops by 5% on average; RSR decreases in both no-XAI (6%) and GenAI conditions (8%). Time pressure has the strongest influence: RAIR and RSR decrease in LIME and drop significantly in GenAI ($p < .01$, $p < .01$, respectively). Participants completed these tasks faster, averaging 20 seconds per task compared to the 30 seconds used in the baseline and payment conditions, despite being given 30 seconds to complete each task. Hence we attribute the performance drop in the time pressure condition to rushing.

We build generalized linear mixed models (GLMM) with Gaussian distributions to investigate how pressure and XAI conditions influence observed RAIR and RSR differences (Appendix C.2.2). All dependent variables are log-transformed, and we consider an effect to be credible if the 97% highest density interval (HDI) excluded zero.

Neither pressure nor XAI support a clear directional effect for RAIR, but both variables are associated with reduced RSR ($\beta = -0.17$, HDI $[-0.33, -0.02]$; $\beta = -0.24$, HDI $[-0.34, -0.07]$). Confirm-

ing our observations that time pressure lowers overall accuracy, the models found pressure condition to have a credible negative association with accuracy ($\beta = -0.03$, HDI $[-.06, -.01]$).

Overall, we find that payment pressure has an observable negative effect on appropriate AI advice reliance and overall accuracy, rejecting our hypothesis that performance pressure will improve appropriate AI reliance. The influence is most pronounced in the GenAI condition. The time pressure condition appears to have stressed participants out, influencing them to rely more on the AI advice, resulting in lower RAIR, RSR, and overall accuracy.

Risk aversion scores have credible negative associations with RAIR but not RSR. AI advisor trust has overall negative associations with appropriate AI reliance. (H2). Our GLMM models find both risk aversion measures, MPQ and the Holt-Laury survey, are associated with reduced RAIR ($\beta = -0.33$, HDI $[-0.61, -0.05]$; $\beta = -0.39$, HDI $[-0.68, -0.10]$) (Appendix C.2.2), but not RSR, despite reporting opposing results: the MPQ personality survey rated participants as risk-loving whereas the Holt-Laury survey rated the same participants as risk-averse, and vice versa. Neither risk score had a credible association with RSR and overall accuracy. We attribute this result to the minimal influence of payment pressure on the dependent variables. Participants' trust in the AI advisor had a credible *negative* association with RAIR ($\beta = -0.89$, HDI $[-1.25, -0.43]$) but a *positive* association with RSR ($\beta = 0.74$, HDI $[0.33, 1.13]$). Neither risk nor trust measures had reliable associations with overall accuracy. From the postsurvey: 10 participants stated they relied on the AI after realizing the first few judgments aligned with the AI advice, indicating trust in the AI advisor, and 7 participants stated they guessed or followed their gut feeling.

Payment pressure increases decision making time (H3). We observe different decision-making behavior when users are operating under payment pressure despite its weak influence measured through RAIR and RSR. In the payment pressure condition, participants on average spent 40% longer making the initial decision and 10% longer considering AI advice before making the final decision, as compared to the baseline pressure condition. This suggests payment pressure could act as a cognitive forcing function, encouraging more analysis before making a decision. Despite the increased time spent on tasks, overall performance was largely unaffected, with a minor increase when no XAI is present and a minor decrease when GenAI is available (Figure 4.2).

XAI can mitigate negative effects of pressure (H4). In contrast to Schemmer et al. (2023), whose work found that LIME improved RAIR and had a minimal effect on RSR, our results show LIME

decreases RAIR compared to no-XAI and reduces RSR in the time pressure condition. Yet LIME synergizes well with payment pressure, improving RAIR over the baseline and maintaining RSR. The RSR change aligns with two participants who stated they relied on the AI significantly and entirely during the time conditions. GenAI explanations improved RAIR and RSR in the baseline condition, but pressure canceled out this effect. However, GenAI marginally improved overall accuracy compared to no-XAI and LIME, indicating the potential for GenAI-style XAI in challenging tasks.

4.4.1 How did participants judge reviews?

Our postsurvey asked participants to describe how they determined whether a review was genuine or deceptive. Several subjects admitted to guessing because they could not determine any relevant features indicating review quality (genuine or deceptive). Twenty-four subjects checked for grammar, typos, and punctuation, and 23 examined the specificity of the reviews. Eight subjects “tried to feel human emotions”, suggesting they associated deceptive reviews with being algorithmically generated, which describes algorithmic aversion for subjective tasks (Castelo et al., 2019). Two subjects stated they were determined not to “lose some bonus”, indicating that the payment pressure condition affected their behavior.

4.5 Discussion

This chapter investigated how environmental pressure, combined with XAI, influences the way crowdworkers complete an AI-assisted decision making task. A large body of literature has focused on improving AI reliance isolated from environmental factors such as pressure, but AI-assisted decision making often occurs under pressure. We demonstrated two simple methods for inducing pressure through payment framed as a loss and limiting time.

4.5.1 AI advice reliance under pressure

There are many types of environmental pressures that an AI-assisted decision maker can be under, and interaction effects can vary widely. When looking at payment pressure, our findings show a subtle effect of how AI advice reliance changes. Without explanations, the pressure conditions had minimal influence on RAIR but negative influence on RSR. With LIME, RAIR improved and RSR remained the same with payment pressure, but RAIR and RSR both decreased with time pressure. With GenAI, RAIR and RSR consistently decreased.

The RAIR decrease in LIME in the baseline pressure condition compared to no-XAI suggests LIME

decreased trust in the AI advice.⁶ It associated individual words with a prediction that may not have made sense (e.g., “is” is associated with being genuine; Figure 4.1b), but users reported associating typos, excessive punctuation, and grammar with their prediction (§4.4.1). Payment pressure appears to mitigate the negative effects of LIME, improving RAIR over the LIME baseline (but not over the no-XAI baseline) without changing RSR. This result points toward the potential for XAI techniques in the same style as LIME to help mitigate or further improve AI-assisted decision making behavior on tasks where humans are close to random.

GenAI has the worst interaction with pressure. RAIR decreases to the level of no-XAI in the payment condition and to the level of LIME in the time condition. We suspect that the payment pressure increased skepticism in AI advice, since natural language explanations can appear generic (e.g., “The review is deceptive because it overly praises the hotel without mentioning any potential downsides”). The time spent considering AI advice increased over the baseline, which can be interpreted as payment pressure influencing people to take more care in their final decision, or reducing their trust in the AI advice in a similar manner to how narcissism has a negative relationship with advice taking (Kausel et al., 2015; O’Reilly and Hall, 2021).

One XAI method can be more effective than another depending on the task or stakeholders (Jiang et al., 2022), and our results show that this is the case even under different pressure conditions. Furthermore, the negative effect of pressure can override the potential benefits of XAI, as seen in the GenAI RAIR decrease.

Time pressure has a largely consistent, negative impact on RAIR, RSR, and overall performance. The behavior change mirrors the phenomenon of “choking,” where a human experiences performance decline in critical high pressure settings, often attributed to anxiety. Participants were anxious about completing the tasks within the 30 second time limit, and rushed to complete them faster. As a result, they relied heavily on the AI advice. One subject stated: “For the timed cases, I ended up relying on the AI’s decision and just went with it. However, for the other parts, I looked out for typos and missing punctuations to detect human reviews.”

Swaroop et al. (2024) observed that slowing down AI overreliors in higher time pressure environments could improve their performance, and since our participants had plenty of time to spare, we expect such a strategy would be helpful in this setting. Emotional regulation strategies can be effective in improving performance when anxiety is high (Balk et al., 2013), which can help alleviate anxiety over running out of time (and hence the possibility of earning a bonus). Methods from the distraction model

⁶Disentangling AI trust and AI reliance behavior is its own challenge (Scharowski et al., 2022; Guo et al., 2024)

may also help: if choking occurs due to cognitive overload (rather than anxiety), an attention shift could refocus the user to pay attention to relevant information and avoid choking (Hardy et al., 2001; Mullen et al., 2005; Eysenck, 2012; Nieuwenhuys and Oudejans, 2012). For example, adding a third color to LIME explanations, unrelated to the prediction, could be a distractor.

4.5.2 Practical Advice

A motivating use case for the work in this chapter is AI-assisted crowdsourced annotation and data elicitation projects in NLP research. In summary, we recommend that GenAI-based explanations be explored for benefits to appropriate AI advice reliance and accuracy. For projects interested in exploring how annotations or elicited data change under different pressure environments, e.g., users may prefer curt dialogue interactions in high pressure environments but elaborate responses otherwise, care should be taken to ensure confirmation bias does not increase AI trust throughout the course of the project. Performance pressure may be a useful tool, but we did not find consistent benefits.

4.6 Future Directions

The work in this chapter has illustrated how environmental stressors and XAI can alter AI-assisted decision making behavior. The mixed effectiveness of incorporating XAI under pressures reinforces the findings of Jiang et al. (2022) and Swaroop et al. (2024) where AI advice or XAI are only effective for some people, some environments, and at certain times. The challenges we observed in determining significant predictive factors of AI advice taking behavior also reflect the difficulty in modeling appropriate AI advice reliance. This calls for building adaptive environments that can be personalized for the user and task to provide appropriate XAI methods, cognitive forcing functions, and other AI advice interventions. Such environments could improve AI-assisted decision making and help elicit more diverse data for training robust NLP models. However, in order to develop these environments, we need a deeper understanding of how environmental factors interact with user characteristics and different XAI methods. Future work should explore other pressures such as competition and multi-tasking, alongside XAI and measures of user personality and behavior, e.g., how they cope with stress. Additional work investigating the influence of performance pressure on tasks where humans have higher than random chance performance should also be investigated, as the literature suggests performance pressure can improve performance, despite our results suggesting minimal influence.

4.7 Limitations

Our experiments in manipulating environmental pressure used an inherently low-stakes, challenging task completed by crowdworkers. As a relatively young area of inquiry, it would be important to investigate how pressure and XAI influence changes between laypeople and experts. We expect different behavior because expertise can influence information seeking behavior (Cathy C. Durham and McLeod, 2000) and buffer stress responses (Matthews et al., 2019).

The literature has yet to find consensus on how people behave in AI-assisted decision making. Our LIME-baseline results contrast with those of Schemmer et al. (2023), despite using a similar interface and pool of crowdworkers. Understanding AI-assisted decision making is a complex and challenging endeavor, and the rapid adoption of AI assistance further motivates research in this area.

4.8 Ethical Considerations

This chapter investigated how pressure and explainable AI influence AI-assisted decision making, and our results show how trust in the AI system plays an important role in agreeing with AI advice. A malicious actor could design a system to increase AI trust in order to persuade others to agree with AI advice, against their better judgment, and overriding beneficial influences of pressure that persuade people to be more skeptical and careful. For example, 10 of our participants stated they found the AI advice agreed with their judgments for the first few instances, leading them to rely on the AI advice more at later stages. Imposing a time limit on the decision could further convince people to rely on the AI advice because relying on the AI is an easy method to cope with time stress.

4.9 Summary

AI assistance is already used in the real world for tasks with high and low risk, and designing adaptive AI assistants that are domain-specific requires understanding the different factors influencing how humans use AI advice. We investigated how pressure influences the use of AI advice. Using deceptive review classification, crowdworkers with little expertise or personal motivation requirements, and two different XAI techniques, we observed complex effects on how pressure influences AI advice usage. Pressure and XAI interactions could both improve and decrease appropriate AI advice. While performance pressure had minor effects, time pressure had a strong negative effect. Our results contribute to the body of literature investigating AI-assisted decision making. We note the relevance of these findings in AI annotation

projects in particular; our work motivates continued research on the effects of pressure on AI assistance in varied environments while taking into account individual differences.

Chapter 5

Conclusion

This thesis presents my key contributions towards creating more realistic environments for studying human-AI interaction and perception within a controlled lab setting. I developed methods and resources that span the research lifecycle and involve different human and human-AI team sizes. The lifecycle begins in Chapter 2 with the data curation process and documentation of how teams interact during collaborative problem solving—a necessary first step to identify how AI can support teams, and how AI should interact with teams. I noted how prior research supports dyadic teams despite varied team designs in the modern workforce and developed a platform for supporting diverse collaborative problem solving research, validating it with a case study and releasing data and code. Next, in Chapter 3, I grounded the theoretical risks and harms of AI literature in an NLP domain where human evaluators can conceptualize potential consequences, revealing how application- and audience-specific environments raise questions on how to address or leverage different model behavior. I created a risk-aware design questionnaire (RADQ) to provide actionable insights into application designs and developed a novel error analysis method (MultiDecoding Analysis) to provide a deeper understanding of model behavior. Finally, in Chapter 4, I show the importance of considering environmental factors, e.g., performance pressure, when studying AI-assisted decision making, illustrating how such factors can override the potential benefits of carefully designed human-AI interactions that were tested in a clean lab environment. I measured the interaction effects of payment and time pressure with two explainable AI techniques designed to mitigate AI overreliance using a neutral binary classification task. My work illustrates how increased realism can be brought into the lab with varying degrees of complexity—and its necessity when designing and deploying AI technology.

5.1 Future Directions

There exist many opportunities for improving the real-world relevance of AI research methods, from data generation to understanding how perceptions of AI influence usage. Future work can directly build on the methods presented in this thesis; new datasets with different team compositions or tasks can be created with `CPS-TaskForge`, risk analyses and error taxonomies compared across different procedural domains and AI models, and pressure influence investigated over tasks with different human performance levels or populations. Looking further afield, other fields of study, such as behavioral economics and game science, can help inspire improved research methods. Chapter 4 drew on the behavioral economics theory of loss aversion; other theories to explore in the context of AI-assisted decision making could be the presence of nudges or social proof. Understanding how humans might make their AI-assisted decisions, even if the decision is irrational, can inform how the AI information is delivered and any interventions required to mitigate undesirable irrational behavior.

Another direction is to improve the role-playing data generation method. Prompting human annotators to role-play is a popular technique designed to enable the collection of synthetic, but realistic, data that would otherwise be challenging to collect for reasons such as expense (e.g., role-playing an expert vs. hiring experts) or availability (e.g., role-playing how a human imagines the ideal robot should respond to different situations). A limitation of the role-playing method is that maintaining the role-played character can be challenging if the human is inexperienced or the environment is not immersive. For example, in the MultiWoZ dataset (Budzianowski et al., 2018), where humans role-played a travel agent scenario, the resulting data lacks sufficient realism, preventing domain transfer to a real-world setting, (Lu et al., 2024b). Game theorists have studied factors that contribute to immersion, one of which is narrative involvement (Calleja, 2011). Perhaps providing a narrative in addition to the character prompt can help players identify with their character and lead to higher immersion, more successful role-playing, and thus more realistic data.

Finally, the fine-grained annotation scheme and error analysis method presented in Chapter 3 can inspire improvements in the fully automatic LLM-as-a-judge evaluation method. This method uses an LLM as a proxy for human evaluation and is scalable due to being fully automatic. Recent improvements to LLM-as-a-judge include asking the LLM-judge to step through a rubric before making its final judgment (Cook et al., 2024). The rubric is automatically created through another prompt. Including domain-specific information during the rubric creation step, e.g., domain-specific errors with higher weighting or desirable behavior that other domains would penalize, could help improve the efficacy of

these rubrics. As AI continues to integrate into daily life, improving our research methods to better simulate and measure real-world interactions will be imperative to ensure our outcomes align with our intentions.

Bibliography

Luis von Ahn. 2013. Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13*, pages 1–2, New York, NY, USA. Association for Computing Machinery.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06*, pages 75–78, New York, NY, USA. Association for Computing Machinery.

Astute Analytica. 2022. Mobilte tower defense games market - industry dynamics, market size, and opportunity forecast to 2030. <https://www.astuteanalytica.com/industry-report/mobile-tower-defense-games-market>. Accessed: 2 June 2024.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hrc map task corpus. *Language and speech*, 34(4):351–366.

Jessica J. Andrews, Tanner Jackson, and Christopher Kurzum. 2019. Collaborative problem solving assessment in an online mathematics task. *ETS Research Report Series*, pages 1–7.

Ars Technica. 2016. Ars Technica. [link].

Robert H. Ashton. 1990. Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification. *Journal of Accounting Research*, 28:148–180.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

- Phillipa Avery, Julian Togelius, Elvis Alistar, and Robert Pieter Van Leeuwen. 2011. Computational intelligence and tower defence games. In *2011 IEEE Congress of Evolutionary Computation (CEC)*, pages 1084–1091. IEEE.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.
- Yannick A. Balk, Marieke A Adriaanse, Denise de Ridder, and Catharine Evers. 2013. Coping under pressure: Employing emotion regulation strategies to enhance performance under pressure. *Journal of Sport and Exercise Psychology*, 35(4):408–418.
- Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *AAAI Conference on Human Computation & Crowdsourcing*.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Edward Beeching, Jilles Dibangoye, Olivier Simonin, and Christian Wolf. 2021. Godot reinforcement learning agents. *arXiv preprint arXiv:2112.03636*.
- Suzanne T. Bell, Shanique G. Brown, Anthony Colaneri, and Neal Outland. 2018. Team composition and the abcs of teamwork. *American psychologist*, 73(4):349.
- Daniel Ben David, Yehezkel S. Resheff, and Talia Tron. 2021. Explainable ai and adoption of financial algorithmic advisors: an experimental study. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 390–400.
- Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward

- mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Vicki M. Bier and Shi-Woei Lin. 2013. On the treatment of uncertainty and variability in making decisions about risk. *Risk Analysis*, 33(10):1899–1907.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Miriam C Buiten. 2019. Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation*, 10(1):41–59.
- Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169.
- Madalina Busuioc. 2021. Accountable artificial intelligence: Holding algorithms to account. *Public administration review*, 81(5):825–836.
- Abhizna Butchibabu, Christopher Sparano-Huiban, Liz Sonenberg, and Julie Shah. 2016. Implicit coordination strategies for effective team communication. *Human Factors*, 58(4):595–610. PMID: 27113991.
- Carrie J. Cai, Samantha Winter, David F. Steiner, Lauren Wilcox, and Michael Terry. 2019. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–24.

- Gordon Calleja. 2011. *In-Game: From Immersion to Incorporation*. The MIT Press.
- Esther Care, Patrick Griffin, Claire Scoular, Nafisa Awwal, and Nathan Zoanetti. 2015. Collaborative problem solving tasks. *Assessment and teaching of 21st century skills: Methods and approach*, pages 85–104.
- Albert V. Carron and Kevin S. Spink. 1993. Team building in an exercise setting. *The Sport Psychologist*, 7(1):8–18.
- Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.
- June M. L. Poon Cathy C. Durham, Edwin A. Locke and Poppy L. McLeod. 2000. Effects of group goals and time pressure on group efficacy, information-seeking strategy, and performance. *Human Performance*, 13(2):115–138.
- Zi-Hang Chen, Li Lin, Chen-Fei Wu, Chao-Feng Li, Rui-Hua Xu, and Ying Sun. 2021. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Communications*, 41(11):1100–1115.
- Karan R. Chhabra, Kathryn I. Pollak, Stephanie J. Lee, Anthony L. Back, Roberta E. Goldman, and James A. Tulskey. 2013. Physician communication styles in initial consultations for hematological cancer. *Patient Education and Counseling*, 93(3):573–578.
- Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing ai-assisted group decision making through llm-powered devil’s advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI ’24*, pages 103–119, New York, NY, USA. Association for Computing Machinery.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021a. All that’s ‘human’ is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021b. All that’s ‘human’ is not gold: Evaluating human evaluation of generated text. In *Annual Meeting of the Association for Computational Linguistics*.

- Jonathan Cook, Tim Rocktäschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. 2024. Ticking all the boxes: Generated checklists improve llm evaluation and generation.
- Christopher C. Corral, Keerthi Shrikar Tatapudi, Verica Buchanan, Lixiao Huang, and Nancy J. Cooke. 2021. Building a synthetic task environment to support artificial social intelligence research. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 65, pages 660–664. SAGE Publications Sage CA: Los Angeles, CA.
- Eduardo Cortes, Vinicius Woloszyn, Arne Binder, Tilo Himmelsbach, Dante Barone, and Sebastian Möller. 2020. An empirical comparison of question classification methods for question answering systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5408–5416, Marseille, France. European Language Resources Association.
- Lawrence A. Crosby, Kenneth R. Evans, and Deborah Cowles. 1990. Relationship quality in services selling: An interpersonal influence perspective. *Journal of Marketing*, 54(3):68–81.
- Shehzaad Dhuliawala, Leonard Adolphs, Rajarshi Das, and Mrinmaya Sachan. 2022. Calibration of machine reading systems at scale. *ArXiv*, abs/2203.10623.
- Steven E. Dilsizian and Eliot L. Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports*, 16:1–8.
- Patricia M. Doney and Joseph P. Cannon. 1997. An examination of the nature of trust in buyer-seller relationships. *Journal of Marketing*, 61(2):35–51.
- Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text.
- Liam Dugan, Daphne Ippolito, Arun Kirubakaran, Sherry Shi, and Chris Callison-Burch. 2022. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *AAAI Conference on Artificial Intelligence*.
- Loïc Duron, Alexis Ducarouge, André Gillibert, Julia Lainé, Christian Allouche, Nicolas Cherel, Zekun Zhang, Nicolas Nitche, Elise Lacave, Aloïs Pourchot, et al. 2021. Assessment of an ai aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. *Radiology*, 300(1):120–129.

- Anna Effenberger, Rhia Singh, Eva Yan, Alane Suhr, and Yoav Artzi. 2021. Analysis of language change in collaborative instruction following. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2803–2811, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Andrea Essl and Stefanie Jaussi. 2017. Choking under time pressure: The influence of deadline-dependent bonus and malus incentive schemes on performance. *Journal of Economic Behavior and Organization*, 133:127–137.
- Michael W. Eysenck. 2012. Anxiety and cognitive performance. *HANDBOOK OF PSYCHOLOGY OF EMOTIONS*, page 87.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.
- Jürgen Fleiß, Elisabeth Bäck, and Stefan Thalmann. 2024. Mitigating algorithm aversion in recruiting: A study on explainable ai for conversational agents. *SIGMIS Database*, 55(1):56–87.
- Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 31–41, San Diego, CA. Association for Computational Linguistics.
- Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake Amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287, Gothenburg, Sweden. Association for Computational Linguistics.
- Jared Freeman, Lixiao Huang, Matt Wood, and Stephen J. Cauffman. 2021. Evaluating artificial social intelligence in an urban search and rescue task environment. In *Aaai fall symposium*, pages 72–84. Springer.

- Shankar Ganesan. 1994. Determinants of long-term orientation in buyer-seller relationships. *Journal of Marketing*, 58(2):1–19.
- Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askill, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1747–1764, New York, NY, USA. Association for Computing Machinery.
- Marisa Garcia. 2024. [link].
- Anna Gausen, Bhaskar Mitra, and Siân Lindley. 2024. A framework for exploring the consequences of ai-mediated enterprise knowledge access and identifying risks to workers. In *ACM Conference on Fairness, Accountability, and Transparency*. ACM.
- Lior Gazit, Ofer Arazy, and Uri Hertz. 2023a. Choosing between human and algorithmic advisors: The role of responsibility sharing. *Computers in Human Behavior: Artificial Humans*, 1(2):100009.
- Lior Gazit, Ofer Arazy, and Uri Hertz. 2023b. Choosing between human and algorithmic advisors: The role of responsibility sharing. *Computers in Human Behavior: Artificial Humans*, 1(2):100009.
- David Gefen, Elena Karahanna, and Detmar W. Straub. 2003. Trust and tam in online shopping: an integrated model. *MIS Q.*, 27(1):51–90.
- Ingo Glöckner, Sven Hartrumpf, and Johannes Leveling. 2007. Logical validation, answer merging and witness selection-a study in multi-stream question answering. In *RIAO*, pages 758–777.
- David Gondek, Adam Lally, Aditya Kalyanpur, J. William Murdock, Pablo Duboue, Lei Zhang, Yue Pan, Zhaoming Qiu, and Chris Welty. 2012. A framework for merging and ranking of answers in deepqa. *IBM J. Res. Dev.*, 56:14.
- Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III. 2024. The impact of explanations on fairness in human-ai decision-making: Protected vs proxy features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 155–180, New York, NY, USA. Association for Computing Machinery.

- Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. 2018. Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2):59–92. PMID: 30497346.
- Ove Grandstrand. 2022. Uncertainty visualization. In JWalter W. Piegorsch, Richard A. Levine, Hao Helen Zhang, and Thomas C. M. Lee, editors, *Computational Statistics in Data Science*, chapter 22, pages 405–421. Wiley, Oxford.
- Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).
- Nina Grgić-Hlača, Claude Castelluccia, and Krishna P. Gummadi. 2022. Taking advice from (dis) similar machines: the impact of human-machine similarity on machine-assisted decision-making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 74–88.
- G. Mark Grimes, Ryan M. Schuetzler, and Justin Scott Giboney. 2021. Mental models and expectation violations in conversational ai interactions. *Decision Support Systems*, 144:113515.
- Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica Hullman. 2024. A decision theoretic framework for measuring ai reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 221–236, New York, NY, USA. Association for Computing Machinery.
- Severin Hacker and Luis von Ahn. 2009. Matchin: eliciting user preferences with an online game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1207–1216, New York, NY, USA. Association for Computing Machinery.
- Nikita Haduong, Alice Gao, and Noah A. Smith. 2023. Risks and NLP design: A case study on procedural document QA. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1248–1269, Toronto, Canada. Association for Computational Linguistics.
- Nikita Haduong, Irene Wang, Bo-Ru Lu, Prithviraj Ammanabrolu, and Noah A. Smith. 2024. CPS-TaskForge: Generating collaborative problem solving environments for diverse communication tasks. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 86–112, Miami, Florida, USA. Association for Computational Linguistics.

- Robin Hale, Jeremy J. Piggott, and Stephen E. Swearer. 2017. Describing and understanding behavioral responses to multiple stressors and multiple stimuli. *Ecology and Evolution*, 7(1):38–47.
- Lew Hardy, Richard Mullen, and Nikki Martin. 2001. Effect of task-relevant cues and state anxiety on motor performance. *Perceptual and motor skills*, 92(3):943–946.
- Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, 70(2):117–133.
- Jennifer Henderson, Maria Kavussanu, Andrew Cooke, and Christopher Ring. 2024. Some pressures are more equal than others: Effects of isolated pressure on performance. *Psychology of Sport and Exercise*, 72:102592.
- Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. 2015. A framework for teachable collaborative problem solving skills. *Assessment and teaching of 21st century skills: Methods and approach*, pages 37–56.
- Randall W. Hill, J. Gratch, Stacy Marsella, Jeff Rickel, W. Swartout, and David R. Traum. 2003. Virtual humans in the mission rehearsal exercise system. *Künstliche Intell.*, 17:5–.
- Martin Hoegl and Hans Georg Gemuenden. 2001. Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization science*, 12(4):435–449.
- John R. Hollenbeck, D. Scott DeRue, and Rick Guzzo. 2004. Bridging the gap between i/o research and hr practice: Improving team composition, team training, and team task design. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, 43(4):353–366.
- Charles A. Holt and Susan K. Laury. 2002. Risk aversion and incentive effects. *The American Economic Review*, 92(5):1644–1655.
- Lixiao Huang, Jared Freeman, Nancy Cooke, Samantha Dubrow, John “JCR” Colonna-Romano, Matt Wood, Verica Buchanan, Stephen Cauffman, and Xiaoyun Yin. 2022. Artificial Social Intelligence for Successful Teams (ASIST) Study 2.
- Susan G Hutchins, Anthony Kendall, and Alex Bordetsky. 2008. Understanding patterns of team collaboration employed to solve unique problems. In *Proceedings of the 13 th International Command and Control Research & Technology Symposium*, pages 17–19.

- Abigail Z. Jacobs and Hanna M. Wallach. 2021. Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- M. Jacobs, M. Pradier, T. McCoy, P. Roy, F. Doshi-Velez, and G. Krzysztof. 2021a. How machine learning recommendations influence clinician treatment selections: example of antidepressant selection. *Translational Psychiatry*, 1:1–9.
- Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021b. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational psychiatry*, 11(1).
- Jinglu Jiang, Surinder Kahai, and Ming Yang. 2022. Who needs explanation and when? juggling explainable ai and user epistemic uncertainty. *International Journal of Human-Computer Studies*, 165:102839.
- Deborah G. Johnson. 2021. Algorithmic accountability in the making. *Social Philosophy and Policy*, 38(2):111–127.
- S Mo Jones-Jang and Yong Jin Park. 2023. How do people react to ai failure? automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, 28(1):zmac029.
- Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. 2024. Understanding trust and reliance development in ai advice: Assessing model accuracy, model explanations, and experiences from previous interactions. *ACM Trans. Interact. Intell. Syst.*, 14(4).
- Edgar E. Kausel, Satoris S. Culbertson, Pedro I. Leiva, Jerel E. Slaughter, and Alexander T. Jackson. 2015. Too arrogant for their own good? why and when narcissists dismiss advice. *Organizational Behavior and Human Decision Processes*, 131:33–50.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

- Harsha Kokel, M. Das, Rakibul Islam, Julia Bonn, Jon Z. Cai, Soham Dan, Anjali Narayan-Chen, Prashant Jayannavar, Janardhan Rao Doppa, J. Hockenmaier, Sriraam Natarajan, Martha Palmer, and Dan Roth. 2022. Human-guided collaborative problem solving: A natural language based framework. *ArXiv*, abs/2207.09566.
- Tom Kontogiannis and Zoe Kossiavelou. 1999. Stress and team performance: principles and challenges for intelligent decision aids. *Safety science*, 33(3):103–128.
- Kundan Krishna, Prakhar Gupta, Sanjana Ramprasad, Byron Wallace, Jeffrey Bigham, and Zachary Lipton. 2023. USB: A unified summarization benchmark across tasks and domains. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8826–8845, Singapore. Association for Computational Linguistics.
- Kundan Krishna, Sanjana Ramprasad, Prakhar Gupta, Byron C. Wallace, Zachary C. Lipton, and Jeffrey P. Bigham. 2024. Genaudit: Fixing factual errors in language model outputs with evidence. *arXiv preprint arXiv:2402.12566*.
- Vivian Lai, Chacha Chen, Qingzi Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: A survey of empirical studies. *ArXiv*, abs/2112.11471.
- 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*, pages 29–38, New York, NY, USA. Association for Computing Machinery.
- Kathryn Ann Lambe, Gary O’Reilly, Brendan D. Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Qual Saf*, 25:808–820.
- Richard N Landers, Kristina N Bauer, and Rachel C Callan. 2017. Gamification of task performance with leaderboards: A goal setting experiment. *Computers in Human Behavior*, 71:508–515.
- Edith Law, Luis von Ahn, and Tom Mitchell. 2009. Search war: a game for improving web search. In *Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP ’09*, page 31, New York, NY, USA. Association for Computing Machinery.
- Benjamin Charles Germain Lee, Doug Downey, Kyle Lo, and Daniel S. Weld. 2023. Limeade: From ai explanations to advice taking. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–29.

- Jiwon Lee. 2015. Analysis of the refinement of shared mental model in science-gifted students' collaborative problem solving process. *Journal of the Korean Association for Research in Science Education*, 35:1049–1062.
- Michelle Seng Ah Lee, L. Floridi, and Alexander Denev. 2020. Innovating with confidence - embedding ai governance and fairness in a financial services risk management framework. *Social Science Research Network*.
- Peter Lee. 2016. [link].
- Gondy Leroy, Stephen Helmreich, and James R Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, 79(6):438–449.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Kuan-Hung Liu, Hung-Chih Chen, Kuan-Ting Lai, Yi-Ying Wu, and Chih-Ping Wei. 2018. Alternative ingredient recommendation: A co-occurrence and ingredient category importance based approach. In *PACIS*.
- Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.
- Vanessa Lopez, Andriy Nikolov, Miriam Fernandez, Marta Sabou, Victoria Uren, and Enrico Motta. 2009. Merging and ranking answers in the semantic web: The wisdom of crowds. In *The Semantic Web*, pages 135–152, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Bo-Ru Lu, Nikita Haduong, Chia-Hsuan Lee, Zeqiu Wu, Hao Cheng, Paul Koester, Jean Utke, Tao Yu, Noah A Smith, and Mari Ostendorf. 2024a. Does collaborative human–lm dialogue generation help information extraction from human–human dialogues? In *First Conference on Language Modeling*.
- Bo-Ru Lu, Nikita Haduong, Chia-Hsuan Lee, Zeqiu Wu, Hao Cheng, Paul Koester, Jean Utke, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2024b. Does Collaborative Human–LM Dialogue Generation

- Help Information Extraction from Human–Human Dialogues? *Proceedings of the 1st Conference on Language Modeling*.
- Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024c. Does more advice help? the effects of second opinions in ai-assisted decision making. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).
- Ioanna Lykourantzou, Angeliki Antoniou, Yannick Naudet, and Steven P. Dow. 2016. Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16*, pages 260–273, New York, NY, USA. Association for Computing Machinery.
- Yingbo Ma, Gloria Ashiya Katuka, Mehmet Celepkolu, and Kristy Elizabeth Boyer. 2023. Automatically Predicting Peer Satisfaction During Collaborative Learning with Linguistic, Acoustic, and Visual Features. *Journal of Educational Data Mining*, 15(2).
- Karl Manheim and Lyric Kaplan. 2019. Artificial intelligence: Risks to privacy and democracy. *Yale Journal of Law and Technology*, 21:106–188.
- Michelle A. Marks, John E. Mathieu, and Stephen J. Zaccaro. 2001. A temporally based framework and taxonomy of team processes. *The Academy of Management Review*, 26(3):356–376.
- Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Hirokazu Masataki, and Yushi Aono. 2018. Role play dialogue aware language models based on conditional hierarchical recurrent encoder-decoder. In *Interspeech*.
- John E. Mathieu, Scott I. Tannenbaum, Jamie S. Donsbach, and George M. Alliger. 2014. A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of management*, 40(1):130–160.
- Gerald Matthews, Ryan W. Wohleber, and Jinchao Lin. 2019. 490stress, skilled performance, and expertise: Overload and beyond. In *The Oxford Handbook of Expertise*. Oxford University Press.
- Carolyn McKay. 2020. Predicting risk in criminal procedure: actuarial tools, algorithms, ai and judicial decision-making. *Current Issues in Criminal Justice*, 32(1):22–39.
- Elisa D Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N Tuch. 2013. Do points, levels and leaderboards harm intrinsic motivation? an empirical analysis of common gamification elements. In *Proceedings of the First International Conference on gameful design, research, and applications*, pages 66–73.

- Kimberly K. Merriman and John R. Deckop. 2007. Loss aversion and variable pay: a motivational perspective. *The International Journal of Human Resource Management*, 18(6):1026–1041.
- Neda Mesbah, Christoph Tauchert, and Peter Buxmann. 2021. Whose advice counts more - man or machine? an experimental investigation of ai-based advice utilization. In *Hawaii International Conference on System Sciences*.
- Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The impact of imperfect xai on human-ai decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–39.
- Henrik Mucha, Sebastian Robert, Ruediger Breitschwerdt, and Michael Fellmann. 2021. Interfaces for explanations in human-ai interaction: proposing a design evaluation approach. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6.
- Richard Mullen, Lew Hardy, and Andrew Tattersall. 2005. The effects of anxiety on motor performance: A test of the conscious processing hypothesis. *Journal of Sport and Exercise Psychology*, 27(2):212–225.
- Larry Neumeister. 2023. [link].
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Arne Nieuwenhuys and Raoul RD Oudejans. 2012. Anxiety and perceptual-motor performance: toward an integrated model of concepts, mechanisms, and processes. *Psychological research*, 76:747–759.
- OECD. 2017. *PISA 2015 collaborative problem-solving framework*. OECD.

- Ami Ooi, Toshiya Iiba, and Kosuke Takano. 2015. Ingredient substitute recommendation for allergy-safe cooking based on food context. *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 444–449.
- Judith Orasanu, Ute Fischer, Yuri Tada, and Norbert Kraft. 2004. Team stress and performance: Implications for long-duration space missions. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 48, pages 552–556. SAGE Publications Sage CA: Los Angeles, CA.
- Charles A. O’Reilly and Nicholas Hall. 2021. Grandiose narcissists and decision making: Impulsive, overconfident, and skeptical of experts-but seldom in doubt. *Personality and Individual Differences*, 168:110280.
- Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.
- Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.
- Luciano Demetrio Santos Pacífico, Larissa F. S. Britto, and Teresa B Ludermir. 2021. Ingredient substitute recommendation based on collaborative filtering and recipe context for automatic allergy-safe recipe generation. *Proceedings of the Brazilian Symposium on Multimedia and the Web*.
- Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)*, UIST ’23, New York, NY, USA. Association for Computing Machinery.
- Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: FAct verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of*

- the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166, Dublin, Ireland. Association for Computational Linguistics.
- Ignacio Pavez, Hugo Gómez, Canlong Liu, and Vicente A. González. 2022. Measuring project team performance: A review and conceptualization. *International Journal of Project Management*, 40(8):951–971.
- Alexandra Paxton, Jennifer M. Roche, Alyssa Ibarra, and Michael K. Tanenhaus. 2021. Predictions of miscommunication in verbal communication during collaborative joint action. *Journal of Speech, Language, and Hearing Research*, 64(2):613–627.
- Mark S. Pfaff. 2012. Negative affect reduces team awareness: The effects of mood and stress on computer-mediated team communication. *Human Factors*, 54(4):560–571.
- Christopher Potts. 2012. Goal-driven answers in the Cards dialogue corpus. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, Somerville, MA. Cascadilla Press.
- Chad A. Proell, Yuepin (Daniel) Zhou, and Mark W. Nelson. 2022. It’s Not Only What You Say ... How Communication Style and Team Culture Affect Audit Issue Follow-Up and Auditor Performance Evaluations. *The Accounting Review*, 97(2):373–395.
- Niels Provos, Panayiotis Mavrommatis, Moheeb Rajab, and Fabian Monrose. 2008. All your iframes point to us.
- Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Question decomposition improves the faithfulness of model-generated reasoning.
- Filippo Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Kim Levin. 2018. Artificial intelligence & human rights: Opportunities & risks.
- Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22.
- Ortwin Renn, Thomas Webler, and Hans Kastenholtz. 1996. *Perception of Uncertainty: Lessons for Risk Management and Communication*, pages 163–181. Springer Netherlands, Dordrecht.

- 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, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.
- Corban G. Rivera, Olivia Lyons, Arielle Summitt, Ayman Fatima, Ji Pak, William Shao, Robert Chalmers, Aryeh Englander, Edward W. Staley, I-Jeng Wang, and Ashley J. Llorens. 2020. Tanksworld: A multi-agent environment for ai safety research. ArXiv preprint.
- Bettina Rockenbach, Abdolkarim Sadrieh, and Barbara Mathauschek. 2007. Teams take the better risks. *Journal of Economic Behavior & Organization*, 63(3):412–422.
- Michelle A. Rodrigues, Si On Yoon, Kathryn B. H. Clancy, and Elizabeth A. L. Stine-Morrow. 2021. What are friends for? the impact of friendship on communicative efficiency and cortisol response during collaborative problem solving among younger and older women. *Journal of Women & Aging*, 33(4):411–427. PMID: 34038325.
- David Ronayne and Daniel Sgroi. 2019. Ignoring good advice. *University of Warwick, Warwick Economics Research Papers Series*, 1150.
- Willibald Ruch, Fabian Gander, Tracey Platt, and Jennifer Hofmann. 2018. Team roles: Their relationships to character strengths and job satisfaction. *The Journal of Positive Psychology*, 13(2):190–199.
- Andreas Rücklé and Iryna Gurevych. 2017. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. In *Proceedings of ACL 2017, System Demonstrations*, pages 19–24, Vancouver, Canada. Association for Computational Linguistics.
- Chantal Savelsbergh, Josette M.P. Gevers, Beatrice I.J.M. Van der Heijden, and Rob F. Poell. 2012. Team role stress: Relationships with team learning and performance in project teams. *Group & organization management*, 37(1):67–100.
- Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and reliance in xai—distinguishing between attitudinal and behavioral measures. In *CHI 2022 TRAIT Workshop on Trust and Reliance in AI-Human Teams*.
- Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's think together! assessing shared mental models, performance, and trust in human-agent teams. *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP).

- Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23*, pages 410–422, New York, NY, USA. Association for Computing Machinery.
- Anuschka Schmitt, Thiemo Wambsganss, Matthias Söllner, and Andreas Janson. 2021. Towards a trust reliance paradox? exploring the gap between perceived trust in and reliance on algorithmic advice. In *ICIS*.
- Omar Shaikh, Caleb Ziems, William Held, Aryan J. Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. *arXiv preprint arXiv:2306.02475*.
- Todd Shore, Theofronia Androulakaki, and Gabriel Skantze. 2018. KTH tangrams: A dataset for research on alignment and conceptual pacts in task-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Helen Smith. 2021. Clinical ai: opacity, accountability, responsibility and liability. *Ai & Society*, 36(2):535–545.
- Janet A. Sniezek and Timothy Buckley. 1995. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes*, 62(2):159–174.
- Angela E.B. Stewart, Arjun Rao, Amanda Michaels, Chen Sun, Nicholas D. Duran, Valerie J. Shute, and Sidney K. D’Mello. 2023. Cpscoach: The design and implementation of intelligent collaborative problem solving feedback. In *Artificial Intelligence in Education - 24th International Conference, AIED 2023, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 695–700, Germany. Springer Science and Business Media Deutschland GmbH.
- Mark Steyvers and Aakriti Kumar. 2023. Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, page 17456916231181102.
- Mike Stoker, Ian Maynard, Joanne Butt, Kate Hays, and Paul Hughes. 2019. The effect of manipulating individual consequences and training demands on experiences of pressure with elite disability shooters. *The sport psychologist*, 33(3):221–227.

- Isabel Straw and Chris Callison-Burch. 2020. Artificial intelligence in mental health and the biases of language based models. *PLOS ONE*, 15(12):1–19.
- Lixin Su, Jiafeng Guo, Yixin Fan, Yanyan Lan, and Xueqi Cheng. 2019. Controlling risk of web question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR’19, pages 115–124, New York, NY, USA. Association for Computing Machinery.
- Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130, Hong Kong, China. Association for Computational Linguistics.
- Chen Sun, Valerie J. Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D’Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers and Education*, 143.
- Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-time tradeoffs in ai-assisted decision making under time pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI ’24, pages 138–154, New York, NY, USA. Association for Computing Machinery.
- Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.
- Samson Tan, Shafiq R. Joty, K. Baxter, Araz Taeihagh, G. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Annual Meeting of the Association for Computational Linguistics*.
- Samson Tan, Araz Taeihagh, and K. Baxter. 2022. The risks of machine learning systems. *ArXiv*, abs/2204.09852.
- Auke Tellegen. 1995/2003. *Multidimensional Personality Questionnaire-276 (MPQ-276) test booklet*. University of Minnesota Press.

- Thora Tenbrink, Elena Andonova, Gesa Schole, and Kenny R. Coventry. 2017. Communicative success in spatial dialogue: The impact of functional features and dialogue strategies. *Language and Speech*, 60(2):318–329. PMID: 28697700.
- Augoustinos Tsiros and Alessandro Palladini. 2020. Towards a human-centric design framework for ai assisted music production. In *New Interfaces for Musical Expression*.
- Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.
- Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, Yanwen Xu, et al. 2024. A survey on human-ai teaming with large pre-trained models. *arXiv preprint arXiv:2403.04931*.
- Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2024. Investigating mysteries of CoT-augmented distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6086, Miami, Florida, USA. Association for Computational Linguistics.
- Xiuxin Wang and Xiufang Du. 2018. Why does advice discounting occur? the combined roles of confidence and trust. *Frontiers in psychology*, 9:2381.
- Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. 2018. Multi-passage machine reading comprehension with cross-passage answer verification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1918–1927, Melbourne, Australia. Association for Computational Linguistics.
- Zhenduo Wang and Qingyao Ai. 2021. Controlling the risk of conversational search via reinforcement learning. *Proceedings of the Web Conference 2021*.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.
- Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *ArXiv*, abs/2304.13835.

- Travis J. Wiltshire, Jonathan E. Butner, and Stephen M. Fiore. 2018. Problem-solving phase transitions during team collaboration. *Cognitive science*, 42(1):129–167.
- Zongda Wu, Shigen Shen, Huxiong Li, Haiping Zhou, and Chenglang Lu. 2021. A basic framework for privacy protection in personalized information retrieval: An effective framework for user privacy protection. *J. Organ. End User Comput.*, 33:1–26.
- Ilan Yaniv and Eli Kleinberger. 2000. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, 83(2):260–281.
- Dan Zakay. 1993. The impact of time perception processes on decision making under time stress. In Ola Svenson and A. John Maule, editors, *Time Pressure and Stress in Human Judgment and Decision Making*, pages 59–72. Springer US, Boston, MA.
- Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, Portorož, Slovenia. European Language Resources Association (ELRA).
- Guanglu Zhang, Leah Chong, Kenneth Kotovsky, and Jonathan Cagan. 2023. Trust in an ai versus a human teammate: The effects of teammate identity and performance on human-ai cooperation. *Computers in Human Behavior*, 139:107536.
- Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "an ideal human" expectations of ai teammates in human-ai teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–25.
- Xiaoge Zhang, Felix T.S. Chan, Chao Yan, and Indranil Bose. 2022. Towards risk-aware artificial intelligence and machine learning systems: An overview. *Decision Support Systems*, 159:113800.
- Hongwei Zhou and Angus G. Forbes. 2022. Data feel: Exploring visual effects in video games to support sensemaking tasks.
- Constantin Zopounidis and Michael Doumpos. 2002. Multi-criteria decision aid in financial decision making: methodologies and literature review. *Journal of Multi-Criteria Decision Analysis*, 11(4-5):167–186.

Appendix A

CPS–TaskForge: Generating Collaborative Problem Solving Environments for Diverse Communication Tasks

— Level 1 Round 1 —

Mundert: no slow :(
Mundert: spam damage?
oobma: sure
Mundert: oh wait
oobma: we got different towers
Mundert: we have different towers
TommyVCT: I guess just yolo it
omar: yeah
Mundert: ok mine only do damage
TommyVCT: I have the one that makes enemies sluggish
TommyVCT: looks like we got a lot of money
omar: mine only do damage too
TommyVCT: oops nevermind we are broke lol
Mundert: easy win
oobma: gogo?
omar: lets go
TommyVCT: gogogo
TommyVCT: it's funny that they went backwards
Mundert: oh it looks like we can kill box with the tree that frightens enemies
Mundert: and the vine one
omar: we probably went overboard lol
Mundert: and area damage would be good with that too
TommyVCT: ez
omar: probably should save money next time to get higher score

— Level 1 Round 2 —

Mundert: wait if we lose do we still get a score
omar: its the same enemies right?
TommyVCT: looks like it's the same
omar: lets have the same setup at the start and nothing after
omar: to save money
Mundert: ok christmas tree and vine killbox?
TommyVCT: I got the same roll of the tools too
Mundert: whatever the cannon was for area damage?
Mundert: spam em
omar: who has the cannons?
oobma: was it the cannon? i only had 1 i thought
oobma: pretty sure it was the plant thing
omar: sorry the catapult
omar: its missing here
Mundert: cannon does area damage
TommyVCT: I'll try to deter the enemies using the diamond
Mundert: so we should use that for a killbox
Mundert: single target is kinda bad for a killbox

(a) Sample conversation from Level 1.

```
<speaker>tjwill</speaker> <chat_text>Full map ones we probably want bottom left </chat_text>
<action>BUY</action> <tower_type>DISCOUNT</tower_type>
<location>(10, 0)</location> <user>ManedWlf</user>
<speaker>tjwill</speaker> <chat_text>If you do a 3x3 grid, empty the center and I'll put an upgrade gem.
</chat_text>
<action>BUY</action> <tower_type>MULTI</tower_type>
<location>(13, 5)</location> <user>schou01</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(0, 14)</location> <user>ManedWlf</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(0, 15)</location> <user>ManedWlf</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(0, 13)</location> <user>ManedWlf</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(1, 13)</location> <user>ManedWlf</user>
<speaker>tjwill</speaker> <chat_text>Then we want a discount tower on the outside, upgrades are Sponive!
</chat_text>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(2, 13)</location> <user>ManedWlf</user>
<action>BUY</action> <tower_type>SUPPORT</tower_type>
<location>(1, 14)</location> <user>tjwill</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(1, 15)</location> <user>ManedWlf</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(2, 15)</location> <user>ManedWlf</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(2, 14)</location> <user>ManedWlf</user>
<action>BUY</action> <tower_type>MAP</tower_type>
<location>(1, 12)</location> <user>ManedWlf</user>
<speaker>schou01</speaker> <chat_text>where do we want to focus our offense? </chat_text>
```

(b) Sample interaction where tjwill suggests placing MAP towers in the bottom left corner of the level in a 3x3 grid, leaving the center empty to place a DISCOUNT tower. ManedWlf proceeds to follow the proposal sending a text message, showing agreement with the proposal through the strategy implementation.

Table A.1: Example conversations and interactions from our CPS-TaskForge pilot study.

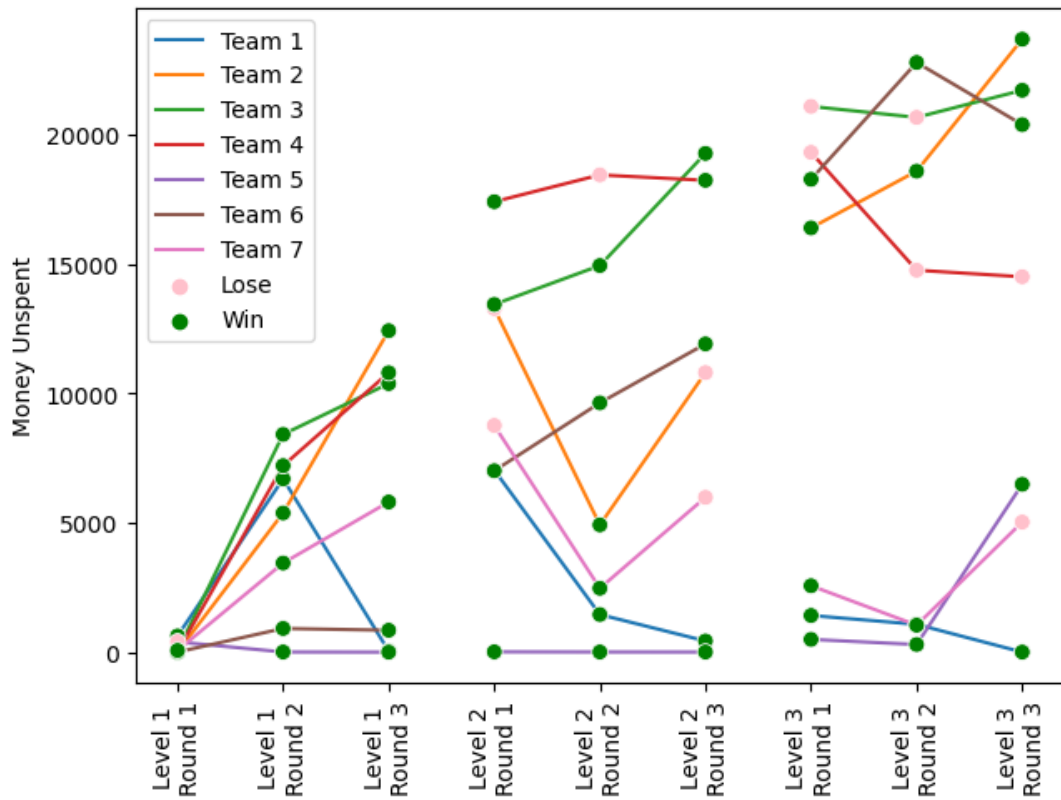


Figure A.1: Money remaining for every team, higher is better. The task goal was to minimize expenditures and still win.

A.1 Survey Questions

The pre-survey collected basic demographic information.

Indicate your age range *

- 18-24
- 25-31
- 32-38
- 39-45
- 46-51
- 52+

What is your highest level of education (or equivalent) completed? *

- Some high school
- High school graduate
- Some college, no degree
- Associates degree
- Certificate program
- Apprenticeship
- Bachelors degree
- Graduate degree
- Other:

What is the area of study of your highest level of education completed? *

For example, economics, physics, literature, foreign language, education, IT networking, not applicable

Your answer _____

Figure A.2: Presuvey part 1

If you are currently in an education program, what is the level of education?

- Associates program
- Certificate program
- Apprenticeship
- Bachelors degree
- Graduate degree
- Other:

What is the area of study of your current education program?
For example, economics, physics, literature, foreign language, education, IT networking

Your answer _____

Race *
Select all that apply

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Other Pacific Islander
- White
- Other: _____

What's your native language? *

Your answer _____

Figure A.3: Presurvey part 2

Describe any proficiency in non-native languages

Your answer _____

Describe your familiarity with the other participants in this study *

For every person, rate them for familiarity from 1-5 where 1 indicates **no familiarity** and 5 indicates *high familiarity*.
 (Optional) Include details explaining your rating.
 Example:

- I don't know anyone.
- I know Mary (2), Jane (3), and don't know anyone else

Your answer _____

How familiar are you with tower defense games? *

1 2 3 4 5

Unfamiliar, don't know what a tower defense is Very familiar, love playing them

How often do you play cooperative games? *

Cooperative game: when everyone on the team works together to achieve a common objective
 ** Does not have to be a video game

Rarely (<2x/mo)

Occasionally (2-4x/mo)

Sometimes (5-10x/mo)

Often (11+x/mo)

Figure A.4: Presurvey part 3

The post-survey contained the Teamwork Quality questionnaire (TWQ, Hoegl and Gemuenden, 2001), VIA Team roles inventory (Ruch et al., 2018), and an open-ended task-specific questionnaire. Both TWQ and VIA used a 7-point Likert scale with options: Strongly Disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, and Strongly Agree.

A.1.1 TWQ

- Communication
 - There was frequent communication within the team
 - The team members communicated mostly directly and personally with each other.
 - There were mediators through whom much communication was conducted.
 - Project-relevant information was shared openly by all team members
 - Important information was kept away from other team members in certain situations.
 - In our team there were conflicts regarding the openness of the information flow.
 - The team members were happy with the timeliness in which they received information from other team members
 - The team members were happy with the precision of the information received from other team members
 - The team members were happy with the usefulness of the information received from other team members

- Coordination
 - The work done on subtasks within the project was closely harmonized.
 - There were clear and fully comprehended goals for subtasks within our team.
 - The goals for subtasks were accepted by all team members.
 - There were conflicting interests in our team regarding subtasks/subgoals.

- Mutual Support
 - The team members helped and supported each other as best they could.
 - If conflicts came up, they were easily and quickly resolved
 - Discussions and controversies were conducted constructively.

- Suggestions and contributions of teammembers were respected
 - Suggestions and contributions of team members were discussed and further developed.
 - Our team was able to reach consensus regarding important issues.
- Effectiveness
 - Going by the results, this project can be regarded as successful.
 - The team was satisfied with the project result.

Open-response questions:

- What went well during the game?
- What went poorly during the game?
- Any notable communication difficulties or frustrations? If they were resolved, how did you resolve them?
- Any notable joyous or satisfactory communications?
- Suppose you played the game again with different maps but the same set of players. What would you change?
- (Optional) Any other comments or complaints about your teamwork or communication?

A.1.2 VIA Team roles

Instructions for participants: for every role, read the description and answer the questions, imagining that you are currently in your ideal team.

- Idea Creator. When working in a team, the creation of new ideas to come up with a solution for a difficult problem or task is essential. Thereby, Idea Creators are people with unconventional ways of coming to solutions and great ideas.
 - In my ideal team, I'm at my best when coming up with ideas.
 - I enjoy creating ideas within my ideal team
 - I am able to be a great idea creator within my ideal team
 - I have a feeling of energized focus when coming up with ideas within my ideal team

- It makes me feel good to create ideas in my ideal team
- Information Gatherer. Information Gatherers search for information, for example on topics as best practices, new trends, potential vendors, competition, and so forth.
 - In my ideal team, I'm at my best when gathering information
 - I enjoy gathering information within my ideal team
 - I am able to be a great information gatherer within my ideal team
 - I have a feeling of energized focus when gathering information within my ideal team
 - It makes me feel good to gather information within my ideal team
- Decision Maker. Decision Makers are processing all the information at hand, integrating it to make the best possible decision and clarifying the goals.
 - In my ideal team, I'm at my best when making decision
 - I enjoy making decisions within my ideal team
 - I am able to be a great decision maker within my ideal team
 - I have a feeling of energized focus when making decisions within my ideal team
 - It makes me feel good to make decisions within my ideal team
- Implementer. Once a team has arrived at a decision on its direction, it needs to implement it. Thereby the Implementer constantly controls the current status and takes measures to work towards the goal.
 - In my ideal team, I'm at my best when implementing goals
 - I enjoy implementing goals within my ideal team
 - I am able to be a great implementer in my ideal team
 - I have a feeling of energized focus when implementing goals in my ideal team
 - It makes me feel good to implement goals in my ideal team
- Influencer. Commonly, the work product of the team needs to be presented by the Influencer for acceptance internally (supervisors, administrators) and/or externally (customers). This is a process of influencing and being persuasive.
 - I'm at my best when representing the work/opinion of the team and convincing others of it

- As a member of my ideal team, I enjoy representing the work/opinion of the team and convincing others of it
 - I am able to be a great influencer in my ideal team
 - I have a feeling of energized focus when representing the work/opinion of my ideal team and when convincing others of it
 - It makes me feel good to represent the work/opinion of my ideal team and convince others of it
- Energizer. In the process of getting work done, Energizers are people that infuse energy into the work and others. Teams without enough energy can fall flat and struggle during times of pressure or prolonged projects that require endurance.
 - In my ideal team, I'm at my best when energizing
 - I enjoy energizing within my ideal team
 - I am able to be a great energizer within my ideal team
 - When I focus on infusing energy into work and others of my ideal team, I feel energized too
 - It makes me feel good to energize within my ideal team
- Relationship Manager. Since the working of a team is a dynamic interplay of people and their relationships, the Relationship Manager helps to run relationships smoothly and to resolve conflicts.
 - In my ideal team, I'm at my best when managing relationships
 - I enjoy managing relationships within my ideal team
 - I am able to be a great relationship manager within my ideal team
 - I have a feeling of energized focus when I manage relationships within my ideal team
 - It makes me feel good to manage relationships within my ideal team

A.2 CPS classification

The CPS skill taxonomy used for classifying utterances in the CPS pilot reproduced from Andrews et al. (2019):

1. Sharing information. Content relevant information communicated during collaboration and includes sharing one's own information, sharing task or resource information, and sharing understanding

2. Maintaining communication. Content irrelevant social communication and includes general off-topic communication, rapport-building communication, and inappropriate communication
3. Establishing shared understanding. Communication in the service of attempting to learn the perspective of others and trying to establish that what has been said is understood.
4. Negotiating. Communication used to express agreement or disagreement and to attempt to resolve conflicts when they arise
5. Exploring and understanding. Actions in the task environment to explore and understand the problem space.
6. Representing and formulating. Actions and communication used to build a coherent mental representation of the problem and formulate hypotheses
7. Planning. Communication used to develop a strategy or plan to solve the problem
8. Executing actions. Actions and communication used in the service of carrying out a plan (e.g., enacting a strategy or communicating to teammates actions one is taking to carry out the plan).
9. Monitoring. Actions and communication used to monitor progress toward the goal and monitor the team's organization

A.2.1 Annotation challenges

Annotating the data for CPS skill using the taxonomy developed by Andrews et al. (2019) was challenging because labels did not have a clear distinction.

For example, consider the following snippet:

- (1) ManedWlf: I have a basic tower with a range of 22, fire rate of 0.8
- (2) ManedWlf: Shall I place a couple close to the castle?
- (3) tjwill: Looks like we've got the same ones to start with, and sounds good!

When ManedWlf describes the basic tower in (1), we can label the utterance for *sharing information* because it is sharing resource information. In (2), a plan is proposed to place some basic towers near the castle, which we can label for *planning*. In (3), we have an observation about both players having the same basic tower. This could be labeled for *sharing information* because tjwill is sharing information

about having access to the same basic tower. It could also be labeled *representing and formulating* because tjwill is building a mental representation about how everyone has the same starting towers.

We defined a few soft rule for classification to help with annotation consistency, but we suggest future work should investigate designing a more complex taxonomy with clearer distinctions between labels.

A few soft rules used when manually classifying CPS skills:

- If a player asks for opinions about placing towers or making upgrades, classify it as Planning.
- If players agree to a plan, classify as Negotiating even if it's just "ok" because it is expressing agreement about a plan proposal.
- If a plan is proposed and another player proposes an alternative or disagrees, classify as Negotiation.
- Representing and formulating is about understanding the efficacy of towers or strategy enacted, e.g., "the blue tower seems to slow enemies down"
- If a player asks someone else to do something, classify as Planning because it is working towards developing the strategy.

A.2.2 Prompt

We tried using automatic annotation with GPT-4, but annotation agreement was only 55%, and developing a CPS classification model with higher accuracy is beyond the scope of this chapter. We list the prompt prefix used for documentation purposes. We used the prompt prefix to classify batches of 6 utterances.

CPS skills list:

```
<skill>Sharing information</skill>. content relevant information  
communicated during collaboration and includes sharing one's own  
information, sharing task or resource information, and sharing  
understanding
```

```
<skill>Maintaining communication</skill>. content irrelevant social  
communication and includes general off-topic communication,  
rapport-building communication, and inappropriate communication
```

<skill>Establishing shared understanding</skill>. communication in the service of attempting to learn the perspective of others and trying to establish that what has been said is understood.

<skill>Negotiating</skill>. communication used to express agreement or disagreement and to attempt to resolve conflicts when they arise

<skill>Representing and formulating</skill>. actions and communication used to build a coherent mental representation of the problem and formulate hypotheses

<skill>Planning</skill>. communication used to develop a strategy or plan to solve the problem

<skill>Executing actions</skill>. actions and communication used in the service of carrying out a plan (e.g., enacting a strategy or communicating to teammates actions one is taking to carry out the plan).

<skill>Monitoring</skill>. actions and communication used to monitor progress toward the goal and monitor the team's organization

You are given a numbered list of inputs. For each input:

Step 1: classify the <chat_text> for one or more <skills> displayed

Step 2: Explain your reasoning in <reason> tags.

Inputs

1. <speaker>ym2552</speaker> <chat_text>It's just when they come in big groups that's worrying, as it seems most towers can only focus on </chat_text>
2. <speaker>schoul</speaker> <chat_text>any chance we can get a buff or discount tower at 9,4?</chat_text>
3. <speaker>jane</speaker> <chat_text>willdo</chat_text>
4. <speaker>paul</speaker> <chat_text>hell, even 1 more turret near the bottom probably would've gotten them all, but we're doing good</chat_text>

Outputs

1. <skill>Representing and formulating</skill>

<reason>The speaker is explaining that when a lot of enemies come at once, they worry the towers will be overwhelmed.</reason>

2. <skill>Planning</skill>

<reason>The speaker is asking another player to place a buff or discount tower at a specific location to further develop the solution</reason>

3. <skill>Executing actions</skill>

<reason>the player is acknowledging a request to act, showing they will execute an action</reason>

4. <skill>Representing and formulating</skill><skill>Maintaining communication</skill>

<reason>the player hypothesizes having one more turret near the bottom would have helped the strategy, then comments the team is doing well to build rapport.</reason>

Inputs

A.3 Potential CPS–TaskForge Tasks

We decided to use the tower defense game genre as the task for CPS–TaskForge after considering several other games.

1. Pandemic TM board game. We found valuable play by forum games that demonstrated the type of multi-turn collaborative communication we hope to see in CPS data. However, one instance of the game takes at minimum 30 minutes to complete, making it challenging to evaluate intermediate task process. The lengthy duration is also a barrier to task repetition within a single study session.
2. Cryptic Crossword puzzles. The cryptic crossword puzzle variant relies on metahints and word-play, making it more accessible than regular crosswords that require trivia knowledge. However, learning the rules is difficult. Participants required 2–3 hours to understand the rules in pilot tests.

The communication during the task was also often short utterances suggesting the solution, with reasoning provided only if teammates requested.

A.4 License

The Godot game engine has an MIT license. The terms for use of our artifacts will be included in our released package.

Appendix B

Human-AI Partnerships: Risks and Harms in NLP

1 1/4 lbs butternut squash diced 1-inch
1 tbsp oil
4 cloves garlic , smashed with the side of a knife
1/4 cup ricotta , I prefer Polly-o
1/4 cup Pecorino Romano , plus optional more for serving
1/4 teaspoon kosher and black pepper , to taste
1/4 teaspoon nutmeg
24 square wonton wrappers
1 large egg , beaten
2 tablespoons salted butter
8 fresh sage leaves , divided

Preheat the oven to 400F. Place butternut, 4 sage leaves and garlic on a sheet pan and toss with 1 tablespoon oil. Season with 1/4 teaspoon salt and pepper, to taste. Roast until tender, about 35 minutes. Transfer to a bowl and mash with a fork until very smooth (a blender would work too). Mix in ricotta and pecorino, season with nutmeg, 1/4 teaspoon salt and black pepper. Place the wonton wrapper on a work surface, brush the edge lightly with egg wash and add 1 tablespoon filling onto the center. Fold over into a triangle and press the edges to seal. Cover with a damp cloth while you make the rest. Chop remaining sage leaves. Place butter and sage in a medium saucepan and melt over low heat. Keep warm over very low heat. Bring a large pot of salted water to a boil. Add half of the ravioli (they are very delicate) and cook until they rise to the surface, about 2 minutes. Use a slotted spoon to remove and add to the pan with the butter. Repeat with the remaining ravioli. Gently toss raviolis with the butter until warm, 1 to 2 minutes. Top with black pepper and serve with additional Pecorino Romano, if desired.

Question: Hello!!! Can you use frozen butternut squash that's already cubed? (Just to save time so I don't have to peel, cube myself?) if so, how would you recommend going about it?

Answer:

1 tablespoon olive oil
2 skinless and boneless chicken breast fillets , halved horizontally to make four fillets
Salt , to season
14 ounces | 400 grams sliced mushrooms
2 teaspoons butter
1 large french shallot , finely chopped (normal shallot for U.S readers)
1/2 cup (about 130ml) champagne (or sparkling white wine)
2/3 cup milk (or heavy / thickened cream)
1 teaspoon of cornstarch (corn flour) – only if using milk

Fresh chopped parsley , to garnish Heat the olive oil in a skillet of pan over medium heat. Season each chicken filet with a pinch of salt. Sear chicken on both sides, for about 3-5 minutes each side (depending on thickness), until golden all over. Transfer chicken to a plate. Fry the mushrooms in the butter and fry for a further 3-5 minutes, or until just beginning to soften. Transfer to the same plate as the chicken. Cover and keep warm. Add the shallot into the pan and cook for 4 minutes, while occasionally stirring. Pour in the champagne; stir well, while scraping any food bits from the bottom of the pan for added flavour. Season with salt and pepper and allow to simmer for about 10 minutes. Add milk (or cream) and simmer for 2-3 minutes. If using milk: mix the cornstarch with a tablespoon of water; pour into the centre of the pan and mix throughout the sauce until slightly thickened. Add the chicken and mushrooms into the sauce along with any cooking juices. Coat in the creamy sauce. Sprinkle with parsley and serve!

Question: Do you have any suggestions for a mushroom substitute? That's a recipe deal breaker for me, otherwise this looks great.

Answer:

Table B.1: Example GPT-3 prompt(s).

Recipe Title	Question	Answer - Human	Answer - GPT3
<i>Pineapple Upside Down Sugar Cookie Cakes</i>	How long can you store the cookies or freeze them?	I haven't frozen them. As for storing, let the cookies cool completely and then place them in an airtight container. If doing a double layer, put a sheet of parchment paper or wax paper between the layers to prevent sticking. Seal the lid, and they can sit on the counter-best eaten within 2-3 days.	The cookies can be stored in an airtight container at room temperature for up to 1 week.
<i>Crock Pot Maple Dijon Chicken Drumsticks</i>	Does anyone know how I could convert this recipe to work in an electric pressure cooker (Instant Pot)?	I would just brown the chicken and set the manual pressure to 15 mins or use the poultry function if you have the 7 in 1 instant pot.	To convert this recipe to work in an electric pressure cooker, you will need to reduce the amount of liquid in the recipe. Maple syrup, balsamic vinegar, and Dijon mustard are all quite thick, so you may need to reduce the amount of each by half. You will also need to reduce the cooking time. Chicken legs usually cook through in about 30 minutes in an electric pressure cooker, so you will need to adjust the cooking time accordingly.
<i>Keto Tomato Tart - Low Carb</i>	First, Happy anniversary, glad you were able to escape and recharge a bit! Second, I'd love to try this base as a dessert; do you think I could do some subs and top it with thinly sliced apple? any specific suggestions? Last, almond flour is sooo expensive, I don't have a Vitamix, I guess I could try my food processor or bkender Have you found a source where it's less than \$12/lb?	Thanks Karen! Yes, you can definitely use this crust for a dessert, I used it for my blueberry crostada in the Fall Recipes Ezine and it works perfectly! Any fruit that you'd use in a pie would work well, including thinly sliced apples. I'm working on a slab pie recipe this week so stay tuned for that. Almond flour IS expensive but for the best results in baking low carb desserts you really have to buy the blanched and finely ground good stuff. I get it on Amazon and buy it in 3 or 5 lb bags to get the best price – around \$10 per pound. Totally worth it! http://amzn.to/1KRiebQ	Yes, I think you could definitely use this base as a dessert crust! I would suggest using a sweeter cheese like mascarpone or cream cheese in the filling, and topping with thinly sliced apples and a drizzle of honey. As for the almond flour, I have found that Trader Joe's carries it for a pretty reasonable price. Hope this helps!

Table B.2: Additional sample questions and answers from the recipes dataset.

B.1 Surveys

Presurvey

How many years of experience do you have **cooking** in a home setting?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

How many years of experience do you have **baking** in a home setting?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

How many years of experience do you have **cooking** in a professional kitchen?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

How many years of experience do you have **baking** in a professional kitchen?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

Figure B.1: Presurvey part 1

Are you **currently** in a culinary arts degree/certificate program or professional apprenticeship?

- Yes
- No

Have you **completed** a culinary arts degree/certificate program or professional apprenticeship?

- Yes
- No

Describe your level of expertise in **cooking**

- Beginner
- Beginner-Intermediate
- Intermediate
- Intermediate-Advanced
- Advanced
- Expert

Describe your level of expertise in **baking**

- Beginner
- Beginner-Intermediate
- Intermediate
- Intermediate-Advanced
- Advanced
- Expert

How often do you **cook**?

- Rarely (0-1 times a week)
- Sometimes (1-3 times a week)
- Often (3+ times a week)

Figure B.2: Presurvey part 2

How often do you **bake**?

- Rarely (0-1 times a week)
- Sometimes (1-3 times a week)
- Often (3+ times a week)

On average, how often do you use recipes published online? Select your most frequent usage.

- 1+ times per week
- 1+ times per month
- 1+ times per year
- Never

For what reasons do you **read** the comments section of a recipe?

- I don't read the comments section
- When I have a question
- No specific reason or purpose when reading the comments section.
- Other

Why do you **post** to the comments section of a recipe?

- I don't make posts.
- When I have a question
- To answer another commenter's question
- To respond to other commenters (but not as an answer to a question)
- To generally publish some writing, such as to show appreciation or excitement about a recipe
- Other

Figure B.3: Presurvey part 3

Indicate your age range

- 18-24
- 25-34
- 35-49
- 50+

What is your highest level of education (or equivalent) completed?

- Some high school
- High school graduate
- Some college, no degree
- Associates degree
- Certificate program
- Apprenticeship
- Bachelors degree
- Graduate degree

(Optional) Anything else you would like to add?

Continue to experiment

Figure B.4: Presurvey part 4

Postsurvey

Do you...

- trust a **human** response more than an AI response
- trust an **AI** response more than a human response
- trust a human and an AI response **equally**
- Other

How important is it to know whether a human or an AI is responding?

- Not important at all
- A little important
- Important
- Very important
- Absolutely necessary

What is your preference for whether a human or an AI is responding?

- Strongly prefer **human** response
- Prefer **human** response
- No preference
- Prefer **AI** response
- Strongly prefer **AI** response

Figure B.5: Postsurvey part 1

Assume you know that an artificial intelligence (AI) is answering questions to recipes (as seen in the study). Do you prefer the AI to...

- Write like another human (not the author) is answering
- Make it obvious a machine is answering
- Write like the recipe author is answering
- Other

Assume you **DON'T** know that an artificial intelligence (AI) is answering questions to recipes (as seen in the study). Do you prefer the AI to...

- Write like another human (not the author) is answering
- Make it obvious a machine is answering
- Write like the recipe author is answering
- Other

[Assume a **HUMAN** is responding] If a question asks for a recommendation that could involve a brand or specific variety [of something], do you prefer...

- Specify exactly the brands and varieties used
- Give generic brand and varieties that could be used
- Other

[Assume an **AI** is responding] If a question asks for a recommendation that could involve a brand or specific variety [of something], do you prefer...

- Specify exactly the brands and varieties used
- Give generic brand and varieties that could be used
- Other

(Optional) Anything else you would like to say?

Figure B.6: Postsurvey part 2

Question/Answer Checklist

Question Looks like a terrific recipe, but what is the oil for? There is nothing in the instructions on how to use it, and the description says it is not needed for the air fryer but helps... do you add it when you add the liquid aminos? Or do you add it to the bottom of the air fryer? Do you coat the rack with it? Please help, thanks!! 😊

Response Oh, sorry! I've updated the recipe card. Thanks for catching that!

Does the response answer the question correctly?

- Yes
 No
 Unsure

How satisfied are you with the response?

Unsatisfied Satisfied

How could the response be improved?

Include more specific information such as...

- Ingredient quantities ⓘ
 Cooking temperatures/times ⓘ
 Nutritional information ⓘ
 Technical details ⓘ
 Expected intermediate or final results ⓘ
 States the recipe has been updated without specifying how in the answer ⓘ
 No improvement possible
 Other

Remove too much extra detail about...

- Ingredient quantities ⓘ
 Cooking temperatures/times ⓘ
 Nutritional information ⓘ
 Technical details ⓘ
 Expected intermediate or final results ⓘ
 Made up details/processes (that may be irrelevant) ⓘ
 No improvement possible
 Other

Change the overall response because it...

- Is too dependent on the specific setting (e.g. ingredients, tools, and skills) of the person following the recipe. ⓘ
 Does not inspire confidence in whether to trust the answer to the question. ⓘ
 No need to change the response.
 Other

Figure B.7: Screen shown when a response answers the question.

How much of the recipe context was required to complete the checklist or determine the correctness of the answer to the question?

- None. I could judge the answer in isolation.
- Required. I needed to refer to the recipe.
- Missing information. I could make some judgments about the answer, but additional information would be helpful.

(Optional) Any additional thoughts?

For example, points you liked or disliked about the answer, more specific details about how the answer is correct or incorrect, or other details about how to improve the answer?

Classify the question. Does the question ask for help with...

- Ingredient alteration [?](#)
- Cooking technique alteration [?](#)
- Preparation/serving [?](#)
- Ingredient clarification [?](#)
- Instruction clarification [?](#)
- Tool clarification [?](#)
- Results clarification [?](#)
- Diagnose failure [?](#)
- Other

Submit response

Figure B.8: Screen shown when annotators state a response does and does not answer the question.

Question/Answer Checklist

Question Looks like a terrific recipe, but what is the oil for? There is nothing in the instructions on how to use it, and the description says it is not needed for the air fryer but helps... do you add it when you add the liquid aminos? Or do you add it to the bottom of the air fryer? Do you coat the rack with it? Please help, thanks!! 😊

Response Oh, sorry! I've updated the recipe card. Thanks for catching that!

Does the response answer the question correctly?

- Yes
- No
- Unsure

Why is it incorrect?

- Describes incorrect ingredient quantities
- Describes incorrect cooking temperatures/times
- Describes incorrect nutritional information
- Refers to incorrect/unrelated ingredients or tools ⓘ
- Describes incorrect or unsafe processes ⓘ
- Describes incorrect or inaccurate intermediate or final results ⓘ
- Is irrelevant/answering a different question ⓘ

Is generic or says it depends on the specific setting (ingredients, tools, skills, etc.) of the person following the recipe.

ⓘ

States the recipe has been updated without specifying how in the answer

ⓘ

Other

Figure B.9: Screen shown when the response does not answer the questions.

Question/Answer Checklist

Question Looks like a terrific recipe, but what is the oil for? There is nothing in the instructions on how to use it, and the description says it is not needed for the air fryer but helps... do you add it when you add the liquid aminos? Or do you add it to the bottom of the air fryer? Do you coat the rack with it? Please help, thanks!! 😊

Response Oh, sorry! I've updated the recipe card. Thanks for catching that!

Does the response answer the question correctly?

- Yes
- No
- Unsure

Why are you unsure? Check all that apply

- Unfamiliar with technique/process
- Unfamiliar with ingredient(s)
- Unfamiliar with the expected result
- Unfamiliar with tools used
- Not enough expertise
- Not enough information in the recipe/question
- Other

Classify the question. Does the question ask for help with...

- Ingredient alteration ⓘ
- Cooking technique alteration ⓘ
- Preparation/serving ⓘ
- Ingredient clarification ⓘ
- Instruction clarification ⓘ
- Tool clarification ⓘ
- Results clarification ⓘ
- Diagnose failure ⓘ
- Other

Submit response

Figure B.10: Screen shown if the annotator is unsure if the response answers the question.

B.2 Survey Results

Crowdworkers Six annotators had 7+ years of home baking and cooking experience, and all but one rated their cooking expertise as intermediate. Baking expertise had a larger range from beginner (3) to intermediate (5). Two annotators requested generic brands when a known machine is responding. Three annotators want a machine to make it obvious that a machine is responding regardless of whether it is known that a machine is responding. There was an even split in trusting human and machine responses equally or trusting a human more.

Experts Two experts had 1–3 years of professional baking and cooking experience, and one expert had 3–7 years of professional cooking experience. Baking experience was rated intermediate by all three, two experts stated they had intermediate-advanced cooking expertise, and one considered themselves an expert. One expert wanted generic brands from a known machine, and one expert did not mind having specific or generic brands. Only one expert wanted a machine to make it obvious that a machine is responding when the respondent is unknown, and another expert wanted similarly when it is known that a machine is responding. Two experts state they trust human responses over an AI response, and one trusts both human and AI responses equally.

Appendix C

How Performance Pressure Influences AI-Assisted Decision Making

C.1 Methods

The natural language explanations for the GenAI XAI condition were generated using ChatGPT with default settings on Jan 20, 2025, with the prefix prompt: In less than 50 words, explain why the following review is a [genuine|deceptive] review.

C.1.1 RAIR and RSR

Relative AI reliance (RAIR) and relative self-reliance (RSR) are defined by Schemmer et al. (2023), which we reproduce below for reference. Subjects complete a prediction task with N instances $x_i \in X$ with ground truth labels $y_i \in Y$. RAIR is the ratio of cases where the human correctly changes their decision to follow AI advice.

$$RAIR = \frac{\sum_{i=0}^N CAIR_i}{\sum_{i=0}^N CA_i} \quad (C.1)$$

where

1. $CAIR$: 1 if the initial judgment disagrees with the ground truth, the AI advice is correct, and the final judgment agrees with the ground truth; 0 otherwise
2. CA : 1 if the initial judgment is incorrect and the AI advice is correct, regardless of the final judgment; 0 otherwise

RSR is the ratio of cases where the human correctly disregards incorrect AI advice.

$$RAIR = \frac{\sum_{i=0}^N CSR_i}{\sum_{i=0}^N IA_i} \quad (C.2)$$

1. *CSR*: 1 if the initial judgment agrees is correct, the AI advice is incorrect, and the final judgment is correct; 0 otherwise
2. *IA*: 1 if the initial judgment is correct and the AI advice is incorrect; 0 otherwise

C.1.2 Presurvey

The presurvey contained demographics questions and the 10-item IPIP adaptation of the Tellegen (1995/2003) Multiple Personality Questionnaire (MPQ). Questions:

1. Indicate your age range
2. Indicate your race
3. Indicate your gender
4. What is your highest level of education completed?
5. What is your native language?
6. Describe your proficiency in other languages
7. How familiar are you with the task of deciding whether a review is genuine or deceptive? (5-point Likert)
8. Rate your level of agreement with the following statements (5-point Likert): (1) I trust artificial intelligence (AI); (2) I use AI to help me with my work

MPQ consists of 10 questions answered on a 4-point Likert scale. Subjects are asked to rate their level of agreement with each statement.

1. I would never go hang gliding or bungee jumping.
2. I would never make a high-risk investment.
3. I avoid dangerous situations.
4. I seek danger.
5. I am willing to try anything once.

6. I do dangerous things.
7. I enjoy being reckless
8. I seek adventure
9. I take risks
10. I do crazy things

C.1.3 Postsurvey

Postsurvey questions:

1. Trust in AI Advisor: Rate your level of agreement with the following statements on a 7-point Likert scale. 1) I think I can trust the AI Advisor; 2) The AI advisor can be trusted to provide reliable support; 3) I trust the AI advisor to keep my best interests in mind; 4) In my opinion, the AI advisor is trustworthy
2. HL survey
3. Describe how you determined whether a review was genuine or deceptive.

The Trust in AI Advisor questions were sourced from Schemmer et al. (2023), who sourced them from Crosby et al., 1990; Doney and Cannon, 1997; Ganesan, 1994 and Gefen et al., 2003.

Holt-Laury Survey

The Holt-Laury survey measures how risk averse a respondent is by asking them to make a decision between pairs of gambles with a “safe” and “risky” choice. For example, in the first question, the participant can select the safe choice of having 1/10 chance to earn \$2 and 9/10 chance of earning \$1.60, or the risky choice of having 1/10 chance to earn \$3.85 and 9/10 chance of earning \$0.10. One of the choices is randomly selected and the gamble played to determine the bonus to the respondent.

Safe choice	Risky choice
1/10 of \$2.00, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10
2/10 of \$2.00, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10
3/10 of \$2.00, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10
4/10 of \$2.00, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10
5/10 of \$2.00, 5/10 of \$1.60	5/10 of \$3.85, 5/10 of \$0.10
6/10 of \$2.00, 4/10 of \$1.60	6/10 of \$3.85, 4/10 of \$0.10
7/10 of \$2.00, 3/10 of \$1.60	7/10 of \$3.85, 3/10 of \$0.10
8/10 of \$2.00, 2/10 of \$1.60	8/10 of \$3.85, 2/10 of \$0.10
9/10 of \$2.00, 1/10 of \$1.60	9/10 of \$3.85, 1/10 of \$0.10
10/10 of \$2.00, 0/10 of \$1.60	10/10 of \$3.85, 0/10 of \$0.10

Table C.1: Holt-Laury survey

C.2 Results

XAI	Pressure	Pre-Advice	Post-Advice
NoXAI	Baseline	0.52	0.55
	Payment	0.54	0.55
	Time	0.51	0.52
	Overall	0.52	0.54
LIME	Baseline	0.52	0.55
	Payment	0.54	0.55
	Time	0.51	0.52
	Overall	0.52	0.54
GenAI	Baseline	0.52	0.55
	Payment	0.54	0.55
	Time	0.51	0.52
	Overall	0.52	0.54

Table C.2: Accuracy before and after receiving AI advice. Accuracy marginally increased after receiving AI advice.

C.2.1 Significance Tests

A Shapiro test for normality indicates our data is not normal. To test for significance, we use the Friedman test with Wilcoxon post-hoc tests for our within-subjects (pressure condition) data and the Kruskal-Wallis and Dunn post-hoc tests for our between-subjects (XAI condition) data. We report uncorrected and Holm-Bonferroni corrected p -values when significant.

Pressure condition tests. For RAIR, when GenAI was present, the difference between baseline and time pressure tasks had p -values of $p = .009$, corrected to $p = .028$. For RSR, when GenAI was present, there was a significant difference between baseline and payment pressures ($p = .008$, corrected $p = .016$), and baseline and time pressures ($p = .003$, corrected $p = .009$).

XAI condition tests. For RSR, GenAI significantly differed from both the baseline no-XAI ($p = 8e-6$, corrected $p = 2.4e-5$) and LIME ($p = 3.4e-4$, corrected $p = .0007$). Similarly, for overall task accuracy, GenAI significantly differed from no-XAI ($p = 1.7e-7$, corrected $p = 5.0e-7$) and LIME ($p = 7e-7$, corrected $p = 1e-6$).

C.2.2 Regression Tables

The generalized linear mixed models using a Gaussian distribution and log-transformed dependent variable results are detailed below. Values in parentheses were transformed back to the original scale.

DV: RAIR	Mean (β)	SD	HDI 3%	HDI 97%	MCSE Mean	MCSE SD	ESS Bulk	ESS Tail	\hat{R}
Sigma	2.101	0.050	2.005	2.192	0.000	0.001	11390.000	6135.000	1.000
Intercept	-1.523	0.225	-1.943	-1.095	0.002	0.003	12491.000	6218.000	1.000
Pressure	-0.107 (0.899)	0.085	-0.269 (0.764)	0.049 (1.05)	0.001	0.001	11686.000	5580.000	1.000
Explanation	-0.07 (0.932)	0.088	-0.232 (0.793)	0.097 (1.102)	0.001	0.001	11010.000	6630.000	1.000
MPQ*	-0.332 (0.717)	0.151	-0.608 (0.544)	-0.047 (0.954)	0.001	0.002	11254.000	6139.000	1.000
Holt-Laury*	-0.391 (0.676)	0.155	-0.682 (0.506)	-0.095 (0.909)	0.001	0.002	12901.000	6782.000	1.000
Advisor Trust*	-0.858 (0.424)	0.218	-1.248 (0.287)	-0.432 (0.649)	0.002	0.002	11377.000	6740.000	1.000
Trust in AI	-0.066 (0.936)	0.064	-0.188 (0.829)	0.052 (1.053)	0.001	0.001	10956.000	6341.000	1.000
Age	-0.121 (0.886)	0.071	-0.249 (0.78)	0.014 (1.014)	0.001	0.001	11665.000	6379.000	1.000
Race	0.015 (1.015)	0.031	-0.042 (0.959)	0.073 (1.076)	0.000	0.000	11200.000	6051.000	1.000
Gender	0.167 (1.182)	0.122	-0.061 (0.941)	0.392 (1.48)	0.001	0.001	11282.000	6049.000	1.000
Education*	0.072 (1.075)	0.027	0.02 (1.02)	0.12 (1.127)	0.000	0.000	11259.000	5936.000	1.000

Table C.3: GLMM results for RAIR. * indicates credible effect (97% HDI excludes 1). Mean and HDI values in parentheses have been transformed back to the original scale.

DV: RSR	Mean (β)	SD	HDI 3%	HDI 97%	MCSE Mean	MCSE SD	ESS Bulk	ESS Tail	\hat{R}
Sigma	2.027	0.048	1.936	2.116	0.000	0.001	12155.000	6525.000	1.000
Intercept	-1.394	0.214	-1.802	-1.003	0.002	0.003	12251.000	6168.000	1.000
Pressure*	-0.169 (0.845)	0.084	-0.331 (0.718)	-0.015 (0.985)	0.001	0.001	12549.000	6463.000	1.000
Explanation*	-0.24 (0.787)	0.085	-0.397 (0.672)	-0.074 (0.929)	0.001	0.001	12539.000	6400.000	1.000
MPQ	0.223 (1.25)	0.143	-0.038 (0.963)	0.491 (1.634)	0.001	0.002	11592.000	6475.000	1.000
Holt-Laury	0.111 (1.117)	0.147	-0.158 (0.854)	0.391 (1.478)	0.001	0.002	12765.000	6687.000	1.000
Advisor Trust*	0.738 (2.092)	0.211	0.33 (1.391)	1.125 (3.08)	0.002	0.002	10998.000	6235.000	1.000
Trust in AI	0.029 (1.029)	0.060	-0.079 (0.924)	0.147 (1.158)	0.001	0.001	11932.000	6689.000	1.000
Age	0.053 (1.054)	0.069	-0.076 (0.927)	0.182 (1.2)	0.001	0.001	11936.000	6174.000	1.000
Race	0.0 (1.0)	0.029	-0.055 (0.946)	0.053 (1.054)	0.000	0.000	12330.000	6358.000	1.000
Gender*	-0.23 (0.795)	0.115	-0.447 (0.64)	-0.018 (0.982)	0.001	0.001	13571.000	6405.000	1.000
Education	-0.019 (0.981)	0.025	-0.067 (0.935)	0.029 (1.029)	0.000	0.000	10931.000	6202.000	1.000

Table C.4: GLMM results for RSR. * indicates credible effect (97% HDI excludes 1). Mean and HDI values in parentheses have been transformed back to the original scale.

DV: Overall Accuracy	Mean (β)	SD	HDI 3%	HDI 97%	MCSE Mean	MCSE SD	ESS Bulk	ESS Tail	\hat{R}
Sigma	0.317	0.008	0.303	0.331	0.000	0.000	11965.000	6325.000	1.000
Intercept	-0.532	0.034	-0.598	-0.472	0.000	0.000	11254.000	6552.000	1.000
Pressure*	-0.032 (0.969)	0.013	-0.056 (0.946)	-0.008 (0.992)	0.000	0.000	12719.000	5529.000	1.000
Explanation*	-0.06 (0.942)	0.013	-0.084 (0.919)	-0.034 (0.967)	0.000	0.000	11629.000	6508.000	1.000
MPQ	-0.011 (0.989)	0.023	-0.052 (0.949)	0.033 (1.034)	0.000	0.000	12045.000	6433.000	1.000
Holt-Laury	-0.018 (0.982)	0.023	-0.06 (0.942)	0.025 (1.025)	0.000	0.000	11983.000	6352.000	1.000
Advisor Trust	-0.013 (0.987)	0.034	-0.077 (0.926)	0.05 (1.051)	0.000	0.000	11390.000	6084.000	1.000
Trust in AI	0.002 (1.002)	0.009	-0.017 (0.983)	0.019 (1.019)	0.000	0.000	11321.000	6503.000	1.000
Age	-0.009 (0.991)	0.011	-0.029 (0.971)	0.012 (1.012)	0.000	0.000	11287.000	6351.000	1.000
Race	-0.003 (0.997)	0.005	-0.012 (0.988)	0.006 (1.006)	0.000	0.000	12220.000	6107.000	1.000
Gender	-0.025 (0.975)	0.018	-0.059 (0.943)	0.011 (1.011)	0.000	0.000	12496.000	6420.000	1.000
Education	0.007 (1.007)	0.004	-0.0 (1.0)	0.015 (1.015)	0.000	0.000	11597.000	6548.000	1.000

Table C.5: GLMM results for overall accuracy. * indicates credible effect (97% HDI excludes 1). Mean and HDI values in parentheses have been transformed back to the original scale.

C.3 Licenses

The Deceptive Opinion Spam Corpus v1.4 (Ott et al., 2011, 2013) is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The outputs from ChatGPT, a large language model from OpenAI, are copyright free.

C.4 AI Use

ChatGPT was used to help with data transformation.