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

Nikita Haduong[♠] **Irene Wang** **Bo-Ru Lu**[♠]
Prithviraj Ammanabrolu[♣] **Noah A. Smith**[♠♢]
[♠]University of Washington [♣]University of California, San Diego [♢]Allen Institute for AI
{qu,nasmith}@cs.washington.edu roylu@washington.edu prithvi@ucsd.edu

Abstract

Teams can outperform individuals; could adding AI teammates further bolster performance of teams solving problems collaboratively? Collaborative problem solving (CPS) research commonly studies teams with two agents (human-human or human-AI), but team research literature finds that, for complex tasks, larger teams are more effective. Progress in studying collaboration with more than two agents, through textual records of team interactions, is hindered by a major data challenge: available CPS corpora are predominantly dyadic, and adapting pre-existing CPS tasks to more agents is non-trivial. We address this data challenge by developing a CPS task generator, CPS-TaskForge, that can produce environments for studying CPS under a wide array of conditions, and releasing a CPS task design checklist grounded in the theoretical PISA 2015 CPS framework to help facilitate the development of CPS corpora with more agents. CPS-TaskForge takes the form of a resource management (tower defense) game, and different CPS tasks can be studied by manipulating game design parameters. We conduct a case study with groups of 3–4 humans to validate production of diverse natural language CPS communication in a game instance produced by CPS-TaskForge. We discuss opportunities for advancing research in CPS (both with human-only and human-AI teams) using different task configurations. We will release data and code.[1]

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 genera-

[1] https://github.com/nhaduong/cps-taskforge

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

tor, 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 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, Zarrieß 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 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., 2021; 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 ac-

	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 ™	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
Maritime Interdiction Operations (Hutchins et al., 2008)	Object identification	3+	Speech
Wiltshire et al. (2018)	NASA Moonbase Alpha Simulation	2	Speech
`CPS-TaskForge` (this work)	Object Identification, Resource Management	1–4+	Text, Speech

Table 1: A sample of collaborative problem solving research. 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 Island™ 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.

tivities in the literature in Table 1 to illustrate gaps in task type and team size between studies with or without data release to the research community.

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 4; 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 towers 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

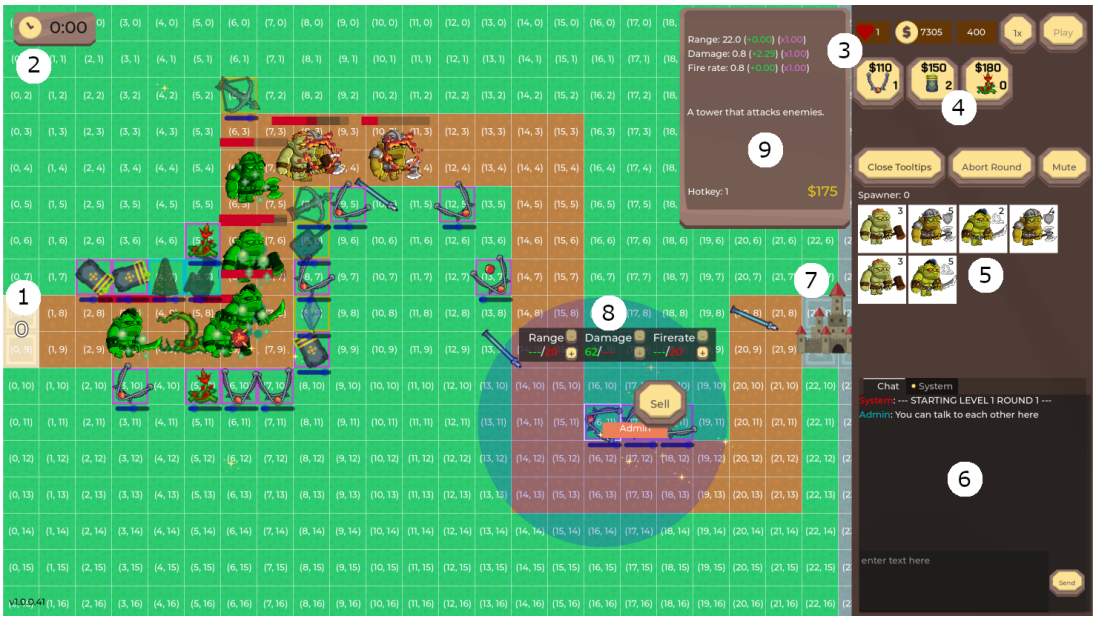


Figure 1: In-game screenshot of a game produced by `CPS-TaskForge`, used in our case study. Enemies spawn from (1) and can only move on the brown path. Towers can only be placed on the green spaces. (2) is the timer used during the *planning* phase, indicating how much time players have to set the board 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.

in cooperative play,[2] 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 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 on the map whose inhabitants can attack the oncoming enemies. The 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.

`CPS-TaskForge` is built on the open-source Godot[3] game engine, and further details of implementation and the tower defense games it produces are available in Appendix A and the documentation of our open-source release.

4 `CPS-✓` : 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-✓`, a design checklist adapted from `PISA2015` (Table 2). We provide default values for `CPS-✓` 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

[2] Bloons TD 6 ™ is a commercial game with a 4-player cooperative mode.

[3] `https://godotengine.org`

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

Table 2: CPS-✓: 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.

different TD games generated by CPS-TaskForge and designed with the help of completing CPS-✓.

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 PentoRef-Take (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.

5 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 reptitions 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 (Appendix Figure 4a; 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 (subsection A.2) 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.

5.1 Data Collection

12 teams of 3–4 people (total 42 individuals) were recruited to participate in a 1.5-hour study[4] 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

[4]Our local IRB approved our study.

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. 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.

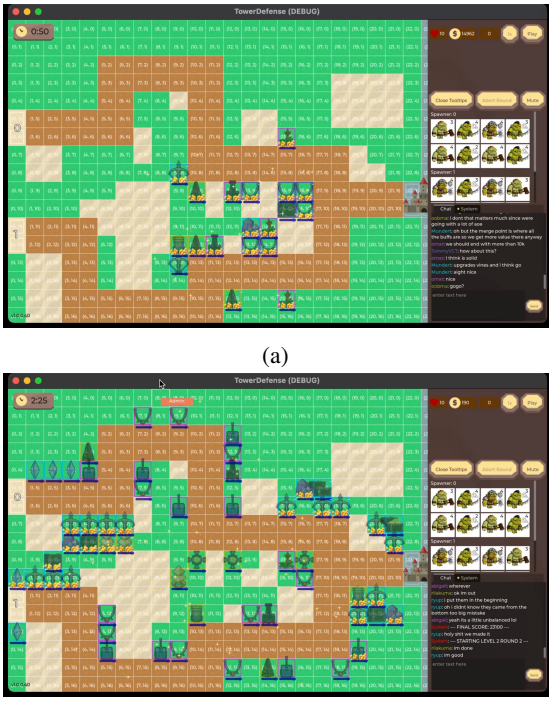
We logged data using XML tags, and the data logged was text communication, score, and tower interaction (upgrading, placing, and selling). The metadata associated with the data was the coordinates of interacted towers, timestamps, and the user. The first 4 teams were used to calibrate game difficulty 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 (Appendix Table 4).

5.2 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 3).[5] 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 Appendix Table 5, 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 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.



(a)

(b)

Figure 2: 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.

5.3 Analysis

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 (Appendix Figure 4b). Figure 2 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 2a received a higher score than Figure 2b. Rounds were repeated three times, allowing teams to op-

[5] We discuss annotation challenges in Appendix Subsection D.1.

Dimension	CPS Skill	Example	Count	Avg. Tokens
Social	Maintaining communication	"haha okay"	222	2.3
	Sharing information	"I have a tower damange 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 3: 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.

timize working solutions–however, teams did not learn to significantly change expenditure behavior, which suggests cautious game behavior (Appendix Figure 5). Teams 1 and 5 appeared to be confused about the task goal, often spending more money across rounds despite winning a previous round.

6 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. Lykourentzou 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., 2021; 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. (2023), Mesbah et al. (2021), and Lu et al. (2024b) 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 success-

fully encouraged more appropriate reliance of AI advice.

7 Conclusion

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. 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`, enables 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. The data generated in our case study reveals different collaborative tasks required to succeed in the overall tower defense task, such as decision making and ensuring teammates have the same understanding of the task.

We will 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.

8 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 implemention 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 competancy 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.

9 Ethical Considerations

The flexibility in designing CPS task environments through `CPS-TaskForge` necessarily places a large responsibility on the designer to design studies appropriate for their target audience or research goal. For example, the imagery used in-game for enemies and towers could be offensive to certain audiences and should be adapted as needed. As with any study in communication, appropriate content filter measures should be in place as required.

The development of generative AI agents as peers that can communicate with humans comes with the risks of the AI agents generating inappropriate content and the concerns of AI replacing humans. Our intentions are that the AI agents can augment human capabilities in more complex problem solving situations, boosting CPS abilities; however, we acknowledge that some problem solving tasks can be simulated and solved through internal or multi-agent collaboration.

Our study was approved by our institution's IRB, and participants were fairly compensated and consented to data sharing with the research community.

References

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 hcrc 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.

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.

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.

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.

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.

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.

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.

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, page 103–119, New York, NY, USA. Association for Computing Machinery.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. 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.

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.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, 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*.

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.

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.

Lior Gazit, Ofer Arazy, and Uri Hertz. 2023. Choosing between human and algorithmic advisors: The role of responsibility sharing. *Computers in Human Behavior: Artificial Humans*, 1(2):100009.

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, page 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.

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.

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, page 1207–1216, New York, NY, USA. Association for Computing Machinery.

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.

Lixiao Huang, Jared Freeman, Nancy Cooke, Samantha Dubrow, John "JCR" Colonna-Romano, Matt Wood, Verica Buchanan, Stephen Caufman, 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.

M. Jacobs, M. Pradier, T. McCoy, P. Roy, F. Doshi-Velez, and G. Krzysztof. 2021. How machine learning recommendations influence clinician treatment selections: example of antidepressant selection. *Translational Psychiatry*, 1:1–9.

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.

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, page 29–38, New York, NY, USA. Association for Computing Machinery.

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.

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.

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 dialogues? *Preprint*, arXiv:2307.07047.

Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024b. Does more advice help? the effects of second opinions in ai-assisted decision making. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Ioanna Lykourentzou, 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, page 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).

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.

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.

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*.

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.

OECD. 2017. *PISA 2015 collaborative problem-solving framework*. OECD.

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.

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.

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.

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.

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.

Chantal Savelsbergh, Josette MP Gevers, Beatrice IJM 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.

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).

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).

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.

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.

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.

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*.

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, page 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, page 75–78, New York, NY, USA. Association for Computing Machinery.

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.

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.

Hongwei Zhou and Angus G. Forbes. 2022. Data feel: Exploring visual effects in video games to support sensemaking tasks. *Preprint*, arXiv:2210.03800.

A `CPS-TaskForge` System Overview

`CPS-TaskForge` is built using the open-source game engine Godot,[6] Nakama,[7] and data collection uses REST API calls to an external server.[8] 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 GodotAgent[9]) 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.

[6] `https://godotengine.org`

[7] `https://heroiclabs.com/nakama/`

[8] The external server we release alongside `CPS-TaskForge` is a Python Flask server.

[9] `https://github.com/Wizzerrd/GodotAgent`

A.1 User Experience

The experience flow is depicted in Figure 3, 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 it allows 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 1; purple color) to indicate who placed which tower. Towers can be placed by clicking a button (Figure 1; 4) or through the assigned hotkey. Tower information is shown in a panel (Figure 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 in-

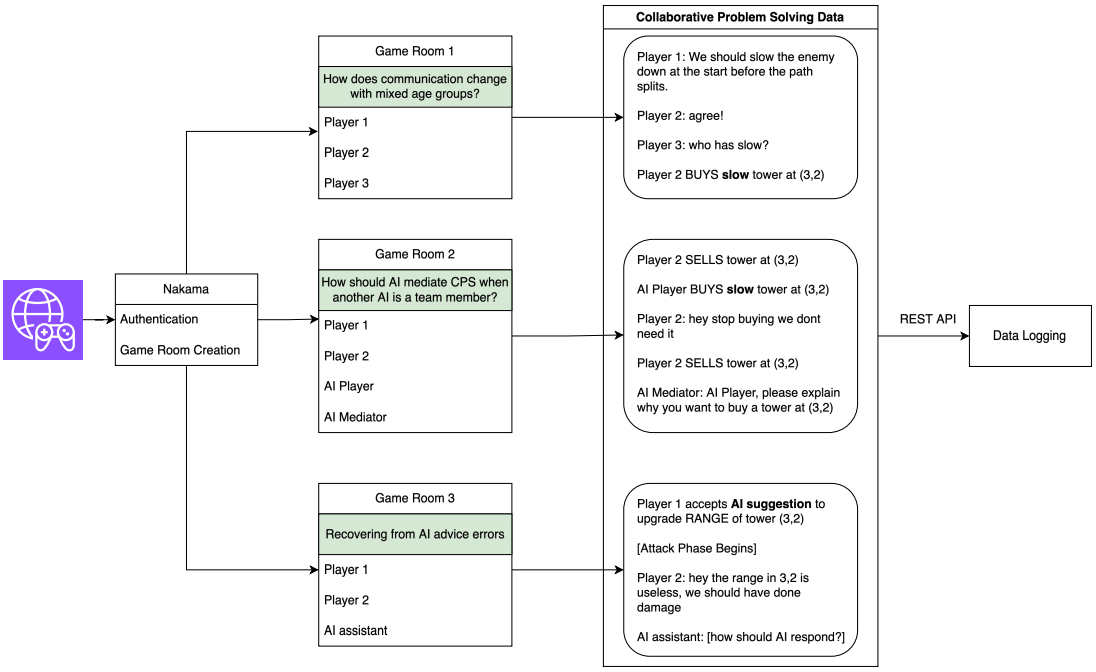


Figure 3: System overview illustrating 3 different research questions that `CPS-TaskForge` supports. Players authenticate through Nakama, join game sessions with different experimental environment designs driven by research questions, and generate CPS data while playing the game. Player interactions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

terface 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 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.

A.2 Tower Defense Designs

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

1. Communication: Voice (bool), push-to-talk (bool), text chat (bool)
2. Description visibility: Tower name (bool), tower description (bool)
3. Number of rounds per level (int)
4. Player resources: Money (shared or individual), health and Score (shared)
5. Interactability during attack phase (bool). Enable this to allow adjusting tower placement and upgrading towers during the dynamic attack phase.
6. 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), multishot (shoots in 4 directions).

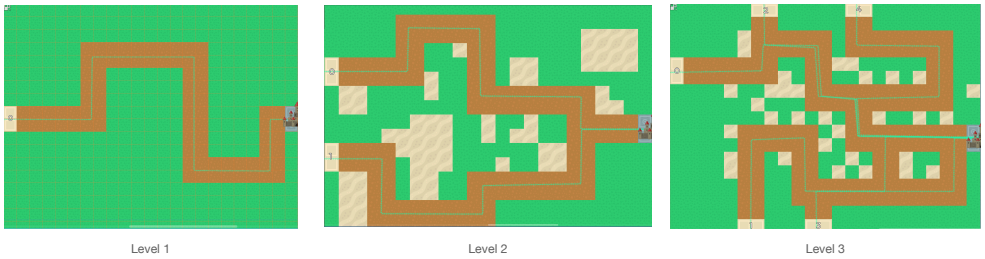
7. Levels: A level design designates how enemies spawn, the enemy movement paths, the location of a base that players defend, terrain for where towers can be placed, starting gold and health, and which towers are available to players.

8. Enemies: There are enemy variants that differ in health, movement speed, point value when destroyed, and money given to players when destroyed.

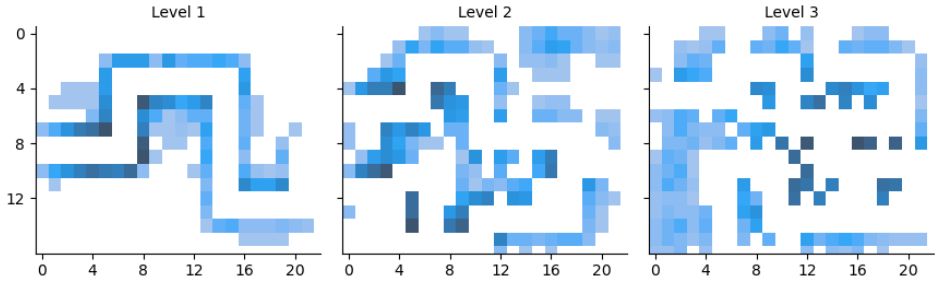
We expect to implement other common game design paradigms such as segmenting the map so players can only place towers on their designated section as the platform matures.

B Case study results

Table 4 describes our case study in the context of other tasks with open data. Figure 2 depicts the levels and tower placing behavior in our case study. Sample conversations are in Table 5.



(a) 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.



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

Figure 4: Game levels and tower deployment in the CPS-TaskForge case study.

	Teams	Participants	Team Size	Tokens	Size	Repetitions	Round Dur.	Study Dur.	Recruitment Platform
TEAMS	63	252	3–4	573k	110k utterances	2	30min	1.5hr	Local
ASIST	64	192	3	—	—	2	15min	3.5hrs	Online, Local
CerealBar	N/A	264	2	325k	24k utterances	N/A	16.5min	—	Crowdworker
PhotoBook	N/A	1,514	2	984k	164.6k utterances	N/A	—	14.2m	Crowdworker
HCRC map task	32	64	2	150k	18hrs	4	—	—	School
PentoRef	63	127	2	216.3k	23k utterances	—	—	—	—
KTHTangrams	42	84	2	68k	11hrs/15k utterances	—	—	15min	Local
Cards	N/A	—	2	282k	45,805 utterances	N/A	8.5min	—	Crowdworker
CPS-TaskForge Pilot	8	35	3–4	8k	1.5k utterances	9	4-6min	1.5hr	Local

Table 4: 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.

— 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 enimies
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
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

(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 Sponsive! </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 5: Example conversations and interactions from our CPS-TaskForge pilot study.

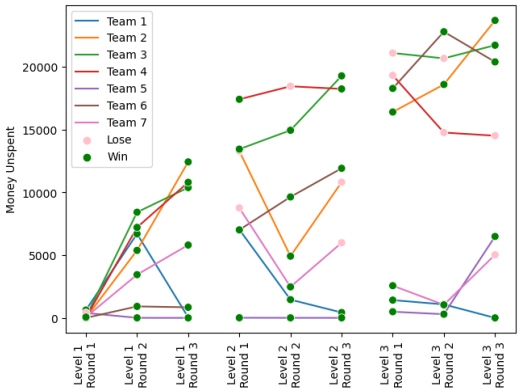


Figure 5: Money remaining for every team, higher is better. The task goal was to minimize expenditures and still win.

C 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

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

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)

The post-survey contained the Teamwork Quality questionnaire (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.

C.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 **d**ifferent maps but the *s*ame set of players. What would you change?
- (Optional) Any other comments or complaints about your teamwork or communication?

C.2 VIA Team roles

Instructions for participants: for every role, read the description and answer the questions, imagining that you are currently in your **i**deal 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

D 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

D.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 en-

acted, 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.

D.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 work. 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<schou1</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
```

E 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 ™ 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 wordplay, 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.

F License

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