I. Full Codebook

TABLE I

THEMES, CODES, DESCRIPTION, AND EXAMPLES USED FOR CODING THE TRANSCRIPTION FOR WORKING GROUP

Theme	*Code*	Description	Example
		Task Requirements	
Scenario and Task	Reconfigurable Scenario	Participant desire for custom and reconfigurable scenarios that allow different modes of research (e.g., optional AI and custom task parameters)	"you're almost going to need a dashboard where you're going to be able to turn on and off a lot of these features."
	Generalizable	Participant desire a task and Minecraft platform that is generalizable and supports many research fields to promote inter-research validity.	"I think it's great to sort of be working toward where we have, we said this before, a common environment where we can share ideas and test stuff."
	Transferable	Participant desire a task that is transferable to richer environments such as the real world (e.g., serves as a proxy for real robots)	"Minecraft... [is] a little bit more complicated, but it's still a grid world at the end of the day, and it's not quite as good as live testing on real robots in the real world."
Team Behavior and Organization	Flexible Organization	Participant support for teams with flexible definitions of roles, interdependence, social structure, and dynamics (e.g., team hierarchy, adversarial teams, multi-objective teams, and multi-team systems)	"That's for example, what we see in police and fire; is there may be flexibility within the organizations so firemen can take over several different roles, and there may be some roles that are interchangeable with police and fire teams"
	Planning-Oriented	Participant desire for AI or teams focused on planning around high-level goals and tasks while independently performing low-level actions	"They're primarily just figuring out a strategy and once the strategy is determined then the low level execution is programmed already."
	Action-Oriented	Participant desire for AI or teams focused on planning around low-level actions under a shared plan	"We don't want to interrupt the people to tell them, hey, you're doing this a little wrong or this a little wrong. Like we want the agent to just jump in and be helping"
Perturbations	Team Perturbation	Participant desire for perturbations in strategy (dynamics), trust (failure), and communication (delay)	"If you are used to your partner and the way they are going to be behaving, but all of a sudden they change their behavior, that could be a type of perturbation."
	Task Perturbation	Participant desire for perturbation in the environment and mission requirements	"You can consider perturbing the environment so that it changes the teamwork"
	Goal Perturbation	Participant desire for perturbation in goals or objectives of the agents	"things that can be captured ... [is] changing the structure of either goals or rewards or priorities of the team and in the mission."

Theme	*Code*	Description	Example
		AI Requirements	
AI Agent Capabilities	Agent Decision-Making	Participant interest in how agents are designed to make decisions	"So one issue is the idea of having the theory of mind, what that means to us is really being able to predict what the team member is going to do."
	Agent Morphology	Participant interest in how an agent is presented (e.g., appearance, embodiment, ...)	"A major interest to us and that's because a big part of our lab studies on the morphology of agents is how the agents are introduced and what they look like."
	Agent Modularity	Participant desire for agents that have a modular design to adapt to various scenarios and models (e.g., large language models)	"... the modularity really helped us there, because then we could just swap out one perception with another perception."
Communication	Multimodal-Based	Participant desire for communication between teammates (human or AI) that uses two or more modes of communication simultaneously or intermittently	"Although we depend heavily on language for our research, we want to sort of adjustability in our methodology where we're not really relying on any one modality."
	Signaling Strategies	Participant desire for AI or teams that can apply various signaling strategies (e.g., affirmation, push/pull, proactive/reactive, implicit/explicit, ...)	"Some of it was directive like if you haven't rescued so many victims in such an amount of time, which is not really a push, and then there was also the "pat on the back" kind of prompts."
	Speech-Based	Participant desire for communication between the teammates (human or AI) that uses natural spoken language	"I definitely appreciate natural spoken language. I think that's the best way for teams to communicate..."
	Action-Based	Participant desire for communication between teammates (human or AI) that uses action with ambigous meaning to communicate intent (e.g., legible motion)	"There's a whole line of work in implicit communication and inferring the intent of the other person as well when that communication is not quite explicit sharing in this condition."
	Text-Based	Participant desire for communication between teammates (human or AI) that uses text or chat features	"There's a chat feature in Minecraft that you can type things and that's very heavily used even with humans between human players to interact with each other."
	Gesture-Based	Participant desire for communication between teammates (human or AI) that uses action with predefined and explicit meaning (e.g., pointing) or other Minecraft gestures or emotes	"But things like pointing, waving to each other, making come hither motions, nodding, yes, shaking your head, no. You know those sorts of actions that could be taken explicitly and then captured... would really enrich the data."

Theme	Code	Description	Example
		Technical Requirements	
Testbed Architecture	API Support	Participant desire for a platform that supports connection to external applications (e.g., SQL) and allows for multiple programming languages (e.g., Docker).	"We know that we build the API's to interface with our models in projects that you know people can develop in whatever they have, their tools, etcetera, and we just build the communication to the Python IBL through some API"
	Scalable	Participant interest in a testbed able to and the feasibility of scaling with the introduction of more users, bigger data, and social structures (i.e. team of teams)	"thinking about the scalability of some of these aspects, right, because one of the big bottlenecks that will hit soon is how do we scale data collection in this space just given how deep models have been taken over on the computer science and AI pieces of it, right?"
	Predefined Terminology	Participant desire for a testbed that comes with a common set of terms, message sets, and definitions to allow easy discussion of components and operation of the architecture	"Write custom code to then get them all to talk nice to each other... the common message structure that then everybody can talk to each other, but you have to talk the same language."
	Operation Modes	Participant interest in a testbed able to run headless (high speed, no visualization) or with visualization	"You often want visualizations for debugging. There's the visualization of the game itself when you're playing... and then there's the fact that you want to run it without any visualizations, headless, so that you can run fast without the overhead of [visualization]. So I think all of those should be supported in general."
	Platform Agnostic	Participant desire for a testbed with the capability to use environments other than Minecraft	"And that's where we found Minecraft, just because it's a game and client that you need to install on your computer, being a bit restrictive."
Data	Data Recording Methods	Participant interest in various methods and scopes of the data collected during task performance	"..we would have access or be able to readily access the data in real-time, which it sounds like that's pretty possible, would be helpful.
	Feedback-Based Methods	Participant desire for data collected from the human providing their feedback about the task or agent specifically in the form of survey or questionare.	"I was wondering if you have any indication from the user when they ask for advice? Do you get feedback from them about whether or not that [advice] was useful?"
Measurement	Agent Measures	Participant desire for predefined measures evaluating agent-level features like effort, self-report, and performance.	"These include programmable agent behavior and post-processing surveys by the participants on their performance as well as AI's performance on the task, either at the end of the task in a debriefing session or rating each advice the AI."
	Team Measures	Participant desire for predefined measures evaluating team-level features like trust, coordination fluency, and communication patterns	"We were quite a fan of the joint activity graphs coming out of HMC as a specification of past interdependencies. So having a strong data representation of the task underlying the simulation is going to be really useful to a lot of researchers."
	Computation Measures	Participant desire for predefined measures evaluating resources used for computation like, speed, memory used and complexity of algorithm	"In general videogame play, you want to compare different approaches with similar computation resources."