This is what will go into ToMCAT Repo

Question-Asking and Plan Inference

Salena Ashton, Loren Rieffer-Champlin, Liang Zhang, Adarsh Pyarelal, Clayton Morrison

1 Introduction

In the SAR scenario for ASIST's Minecraft virtual environment, teams of human players engage in cooperative behavior to search for, stabilize, and rescue victims within a collapsed building. As human players verbalize plans, make suggestions, or tell each other what to do, they also ask questions that can infer hidden goals or intentions. Teammates reduce their individual knowledge asymmetry by asking and answering questions. Using Theory of Mind (ToM), [1], we will investigate how uttered questions can infer another person's goal or intention.

This investigation will guide future research into knowledge engineering and representation of tasks, goals, and hidden intentions. While simple human goals may be represented with classical AI planning, complex goals that have constraints or multiple levels of abstraction may be best represented by a *hierarchical task network* (HTN), which is a tree of possible plans.[2]

We will investigate the following research questions for Study-3:

1. How do spoken questions reveal another person's plan or intent?
2. Can people listen to spoken questions and accurately decide if the other person's plan is simple, sequential, or hierarchical? Can the distinction between such structures improve predictive performance for Artificial Social Intelligence (ASI) agent intervention?

Previous research into question-*answering, not asking,* centered on optimization because researchers assumed that people prefer concise questions and answers. However, people do not engage in dialog using question or answer sets. They may ask open-ended questions, meander in speech without purpose, and use indirect speech acts to express mutual goals or build rapport with each other. The current literature regarding ToM and question-answering is sparse but even more so for question *asking*.

Hawkins and Goodman connect question-asking and intention *because* of the scarcity of "empirical evidence about how social context affects the questioner's choice." They redefined the meaning of a question as "the interpretation and response of a hidden (or uttered) goal, to be discerned by another person, typically a dialog partner" [**?**]. Hawkins and Goodman describe speech acts and question-answer dialog as a form of *information asymmetry*: A questioner has a goal but needs information while an answerer has information but does not know the questioner's goal. The type of questions then asked will depend on context, social inference, and signals in a dialog setting. The significance of their work is in the decoupling of the inferred goal from the explicit meaning of the question to model context and avoid assumptions. [3]

[1] The capacity to infer another's thoughts, feelings, beliefs or intentions.

[2] Technically, the plans produced by HTN planners can also be represented with flat lists - however, in this section, we use the term 'plan' to refer to the actual 'plan tree' that contains the task decompositions as well, rather than just the plan alone.

[3] Their work was limited to epistemic questions and cooperative behavior.

Deciphering a person's goal or intention from their answer to a question, instead of the question itself, may be another way to understand intent. While Hawkins and Goodman define questions as hidden goals or intentions, Mehdi Alaima et al define the act of asking questions as 'the providing of information or knowledge to reinforce knowledge one way or another," independently paralleling the definition given by Hawkins and Goodman. "When information is missing, or contradicts what one knows, a knowledge goal will arise, often leading to the generation of questions. The person is then made aware of the information needs, and motivated to formulate a question to obtain the missing knowledge," [**?**]. Building on the claims of Hawkins and Goodman, and Alaima et al, we investigate question-asking within the SAR scenario of ASIST's Minecraft environment.

2 Approach

We assume that questions have hidden goals and infer plans. As teams ask more questions of each other, human team ToM converges toward cooperative behavior. We will investigate whether question-asking is associated with *team planning*, defined as a set of goals, strategies, or tasks that are executed. We define *coordination* as behaviors and utterances to create a common plan or strategy[4]. We define *cooperation* as team behaviors that implement an already-agreed up on plan.

To capture hidden goals, inferred plans, and patterns that may represent human ToM, we will annotate six ASIST Study 3 Spiral 2 pilot video observations and six HSR ASIST Study 3 videos released between March 29 and May 5, for a total of at least twelve videos. These videos are of three distinct missions for each team. Due to the expensive costs of taxonomy label development with strict adherence to grounded theory methodology, this stage of the experiment is limited to no less than twelve videos.

Two human annotators will code all uttered questions between teammates within the Minecraft SAR scenario. We use the qualitative coding procedure known as Grounded Theory, as defined by **?**.

More specifically, and as defined by **?**, we will use a Grounded Theory Process Coding for a state or action across some interval of time. These *grounded-in-data* labels are known as *concept-level* labels, which are the smallest pieces of data that encode a question-asking phenomena of interest. We also use this coding methodology to investigate the connectivity and causality of each concept label to discover possible relationships between presence or absence of team actions, interactions, conditions, and consequences of question-asking. Densely-connected concept labels suggest subcategories and categories. Sparsely-connected labels will not be discarded; they will be used to consider variability within patterns and categories that emerge. In cases where questions have co-reference or other contextual dependencies, only that direct dependency will be coded for local semantic meaning.

We make the following considerations when creating codes:

- Frequency will not dictate importance, causality, or connectivity of a concept
- Each question will have at least one annotation and up to four annotations:
  - Primitive actions (ground truth). Ex: breakRubble, requestStabilizedVictimCarry.
  - Abstraction Levels of actions (of primitive actions) Respective examples: respondRubbleRequest or createVictimAccess, collaborateStabilizedVictim
- Labels will be stemmed and minimally normalized
- Capturing the phenomena of question-asking across time, between any subset of a team, between the same team across the two different missions.

To avoid annotator and researcher biases and any *a priori* belief on which team ToM strategies may be used, concept and category labels are not pre-determined. Inter-annotator agreement must reach a Kappa Score of 80% or higher. This also gives a more solid, grounded analytical meaning to any emergent categories.

[4] Note that this is distinct from the mathematical definition of coordination proposed in **??**

After the development of labels and taxonomy, the investigation of team ToM and question-asking will scale for additional videos. When all concepts, subcategories, categories can reasonably explain the phenomena of the video observations, one or two super-categories, *theories of team plan*, will emerge. We currently assume that a theory of team plan would have greater predictive power and ToM inference potential.

3 Evaluation

Because of the small sample size of this investigation, we will not perform a quantitative evaluation at this time. Instead, we will perform a qualitative investigation of word frequencies, clustering patterns, and correlation of annotator-generated labels through data visualization. Below is a list of possible visualizations we may consider:

- Connectivity of concept-level labels: radial diagrams, arc diagrams, matrix diagrams or graph networks
- Frequency patterns of words or concept labels (normalized, word count / total number of words in that question): scatterplots or histograms
- Correlation of words and labels with time: time series, scatterplots.
- Concept-level subcategorization(s): clustering, PCA (concept labels possibly projected onto sub-categorical spaces), or hierarchical visualizations.

Such visualizations, based on twelve videos, will lead to further insight through this investigation. Future measures may include Mann-Whitney U-Tests, t-tests (only if we annotate a large-enough sample), precision and recall of the concept-level and category patterns to describe the generalizability for real data with no ASI interventions, generalizability for real data with ASI interventions, and the variance of patterns in label categories. Another possible measure, for future research, would be the F1 score to explain how well these labels describe observations without ASI interventions, when compared to high-intervention observations. This future investigation would address measure ASI-M5: Coordinative Communications to measure teamwork, include additional video observations for real data in Study 3, and continue our investigation of whether human plans and ToM are best represented by classical planning or HTN planning.

4 PROMISES MADE

- investigate how uttered questions can infer another person's goal or intention.
- annotate six ASIST Study 3 Spiral 2 pilot video observations and six HSR ASIST Study 3 videos released between March 29 and May 5, for a total of at least twelve videos. Two human annotators
- Strict adherence to Grounded Theory (bottomup/ process and causal coding)
- Kappa $> 80\%$
- emergent categories; 1 - 2 super theories will emerge, which we assume to mean Team ToM
- Preliminary Analysis HOW Team TOM maps to AI PLanning– due to small sample size:
  - qualitative investigation of word frequencies, clustering and/ correlation
  - qualitative investigation through visualization
  - scatterplots or histograms or hierarchical visualizations
  - information across time

5 Results

5.1 Label Development and Taxonomy

We investigated the connection between question-asking and hidden goals of the question-asker with strict qualitative coding approach known as Grounded Theory, ensuring that our annotations were not biased by prior research, predetermined labels, or theories about player strategy and ToM. This data-driven, robust method captures player intention and team strategies of interest and naturally lends itself to further quantitative analysis.

Two independent annotators[5] viewed six study-3 pilot videos using no set of labels. Videos were selected from no-human intervention videos, at random, and three teams were chosen as a sequence (team 217, 218 and 219). The continuous disambiguation of unsupervised labels led to extensive documentation of verb, noun and modifier use agreement for the annotation of six different study-3 real data videos[6], which then lead to Theoretical Saturation[7] At saturation, our unsupervised labels yield 21 verb labels, 24 noun labels, 25 modifier labels. Without replacement, this yields 12,600 possible label combinations, yet Ashton demonstrated theoretical saturation through the independent construction of less than 100 labels from two autonomous annotators.

5.2 Data Cleaning and Annotator Agreement

We achieved an unweighted Cohen Kappa[8] agreement of 0.892. Using the disambiguation documentation and Merriam-Webster's Dictionary to meticulously settle any annotator label disagreement, Ashton declared the question labels as 'agree' or 'disagree'. When annotator intention and the disambiguation documentation did not clarify the agreement or disagreement of labels, Ashton declared it as 'disagreement'.

5.3 Emergent Categories, Team Theories, and Preliminary Patterns

The 100 labels that resulted from two annotators were then analyzed to discover emergent categories and preliminary patterns. Frequency of labels *do not* dictate the importance of a label but do alter probabilistic analysis. The data were minimally lemmatized and yielded the following results: Joint probabilities of questionLabel[9],

abstractLabel[10],

The data were minimally lemmatized and aspects of label concept granularities were resolved using conditional probability. For example, the conditional probabilities of questions, given a player intention would change if "critical" and "victim" were labeled as "victim". Therefore, these and similar granularities were maintained. Other conditional probabilities, such as "location" and "room" did not significantly change, so after the data were cleaned, a global replacement of "room" to "location", and similar replacements were made. Global placement after post processing is not reflected in the Kappa score in order to maintain raw data and researcher integrity.

We did not regard label frequency as criteria for label usefulness. Instead, from these emergent categories, we chose specific words that optimized intention or belief among player interaction[11].

[5] Ashton and Kim

[6] Ashton and Reiffer-Champlin

[7] When unsupervised label development from fully-autonomous annotators allows for additional labels, yet very few are created. It is this point that shows how the existing label schema fully capture the events, data and phenomena observed.

[8] In order to adhere to the robust procedure of Grounded Theory, the technical calculation of Kappa's score would include the probability of *all possible probabilities of all labels*, which would approach zero. While it sounds trivial, it must be mentioned that the reason Grounded Theory is robust in its unsupervised labeling approach is because annotators have complete autonomy. Because the sample size is small and only two annotators worked on the real data, there is no current justification for using Scott's pi or a weighted Kappa. This will change as investigation and further research scale.

[9] Code representation of the question utterance and no further context except for co-reference resolution. Some questions were spoken as statements with inflection. Other questions were actually demands shrouded with politeness. All three types of questions were coded for this investigation.

[10] Code representation of player intention. This was captured by taking the question into context. For example, if the question were represented by 'requestBreakRubble', the intent behind a request could have been 'collaborateCriticalWake' or 'navigateLocation'. AbstractLabel is not the representation of *why* a question had been asked; it is the *intention* of the player who asked the question.

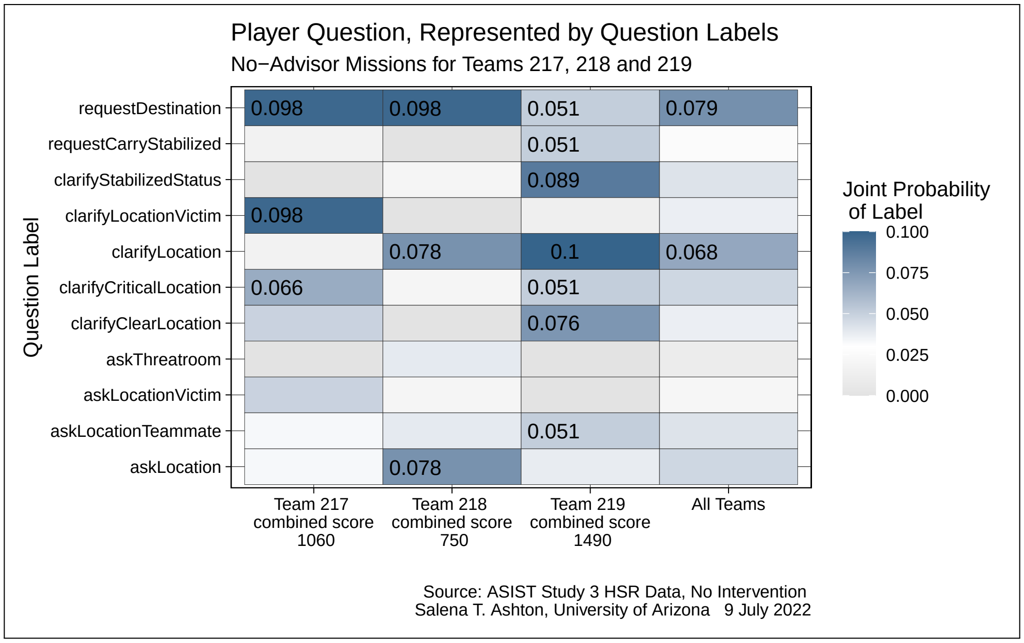
[11] For example: *ask*, *request*, *suggest*, and *clarify* are all actions that show the goal of obtaining information, but clarify asks for additional

5.4 Emergent Categories of Player Intention

From more than 400 unconstrained labels developed in stage 1 annotation, we discovered recurrent components from which goals or intentions emerged:

- Less Team Collaboration: Talking *at or to* a teammate and not *with* a teammate: direct, suggest, etc.
- More Team Collaboration: Talking *with* a teammate: ask, request, answer, clarify, etc.
- Intention toward position: location, navigate, destination, room, etc.
- Prioritizing an idea: plan, suggest, request, collaborate, critical, victim, etc.
- Question utterances that were actually demands, statements that were questions with inflection, and other nuanced utterances: suggest, tell, direct, request; context and explanations included in annotation.
- Autonomy in any other label creation; this resulted in marker, mission, wake, etc.

information needed before acting upon a goal or task, request is an initiation of commitment to another player with the optional accept/ reject response that suggest would not imply, and ask is a general desire for information.

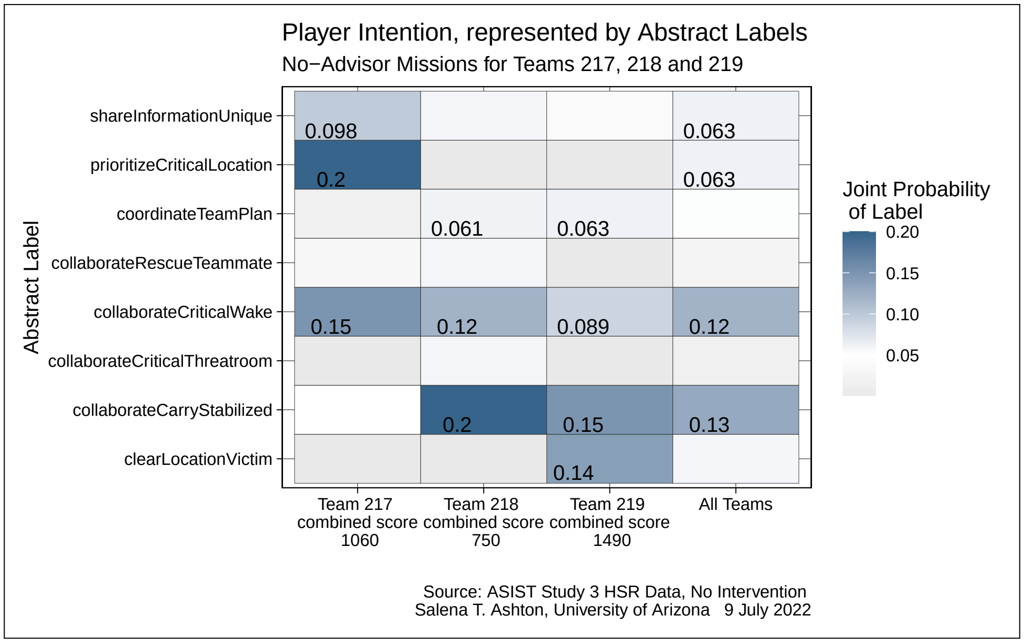


Source: ASIST Study 3 HSR Data, No Intervention
Salena T. Ashton, University of Arizona 9 July 2022

Figure 1: Player question utterances, represented by question labels and displayed by each team and the most probable question labels.

Most common questions: requests for teammates to come; clarification of a victim's status for transport (shown by the word 'stabilized' and not by 'victim' (an unstabilized victim); clarification of any victim's location and common questions about location.

Notice that team 219, who scored the highest of the three, asked more questions to clarify communication with each other. Previous assumptions about players who prioritized the clearing of rubble, navigation strategy or dividing up tasks is not supported in this preliminary investigation.


Player Intention, represented by Abstract Labels

No−Advisor Missions for Teams 217, 218 and 219

Source: ASIST Study 3 HSR Data, No Intervention
Salena T. Ashton, University of Arizona 9 July 2022

Figure 2: Player intention, represented by abstract labels and displayed by each team and the most probable abstract label.

The most common intentions, expressed by questions and other utterances, visual cues, context, and player interaction (represented as an abstract label) are collaborating with teammates about waking critical victims and carrying stabilized victims to the proper sick bay.

Team 219, who scored some of the highest combined scores in many of the observations, not just the ones annotated, compared to 218, who scored some of the lowest in the data, intended to collaborate on carrying stabilized victims to the proper sick bay. The difference in score is found in more granular actions: while team 219 often requested for collaborative tasks, team 218 would more often demand or tell others what tasks to perform (as seen in figure 1).

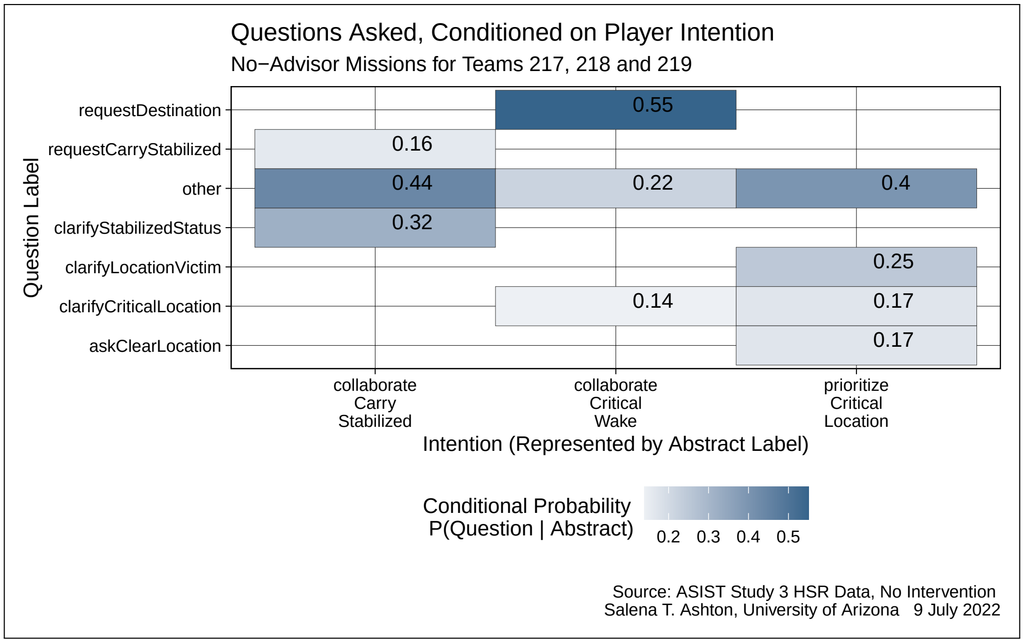


Figure 3: Conditional probability of questions asked, given the player's intention.

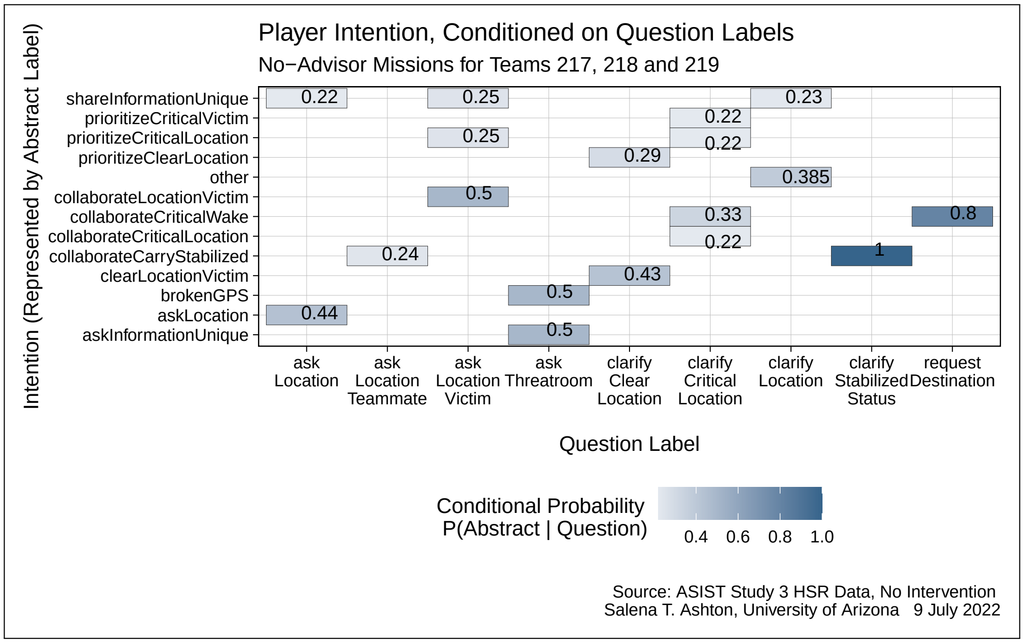
It is trivial to see the request of teammates to come to a particular location in order to wake up a critical, and so it is no surprise to see this is the most common reason for 'requestDestination' questions to be asked. It is far more interesting to see (figure 4) that 'requestDestination' is more than 80% to be asked only for the intent of waking a victim.

What is most interesting is the variation of questions that can be asked, given player intention. Notice that more than 40% of questions that could be asked are marked 'other.' To include each instance would make a noisy and sparse visual. It is this variation that matters: players will not necessarily ask expected questions (or any question at all) just because they have intention.

The most likely questions that *are vocalized* have one aspect in common: they all have verbs that are collaborative in nature: request, clarify, collaborate. Verbs that are less collaborative in nature, (direct, suggest, inform) do not show up as the most common questions.

This suggests that when players feel like they are part of the team, they are more likely to ask questions that have a give-and-take nature.

5.4.1 Questions and Intentions


Player Intention, Conditioned on Question Labels

No−Advisor Missions for Teams 217, 218 and 219

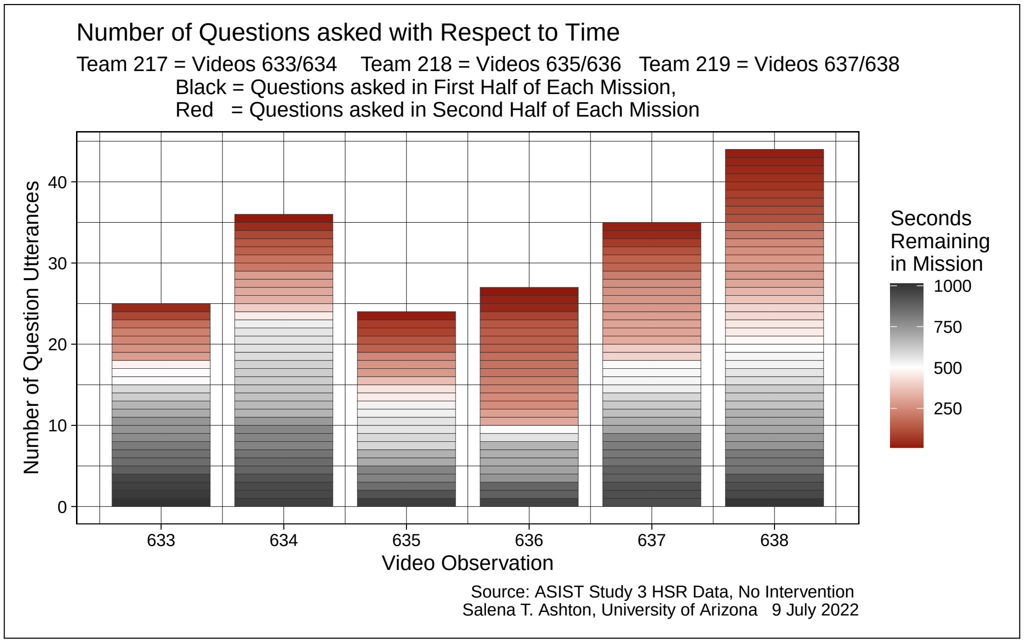
Source: ASIST Study 3 HSR Data, No Intervention
Salena T. Ashton, University of Arizona 9 July 2022

Figure 4: Conditional probability of player intention, given the question that is asked.

I remind the reader that we did not take label frequency into account when determining emergent categories and behavior patterns. To do so would bias the researcher and annotators. Here, as in other figures, the probability of a question that clarifies the stabilized victiim's status has a theoretical probability[12] approaching 100% intent to collaborate with others on where to deliver that victim.

Players were most likely to ask for unique information from the engineer when they wanted to prioritize locations or learn about threatrooms. When the GPS systems failed, more than half of the time players asked about threatrooms than they did victims, rubble, or other events.

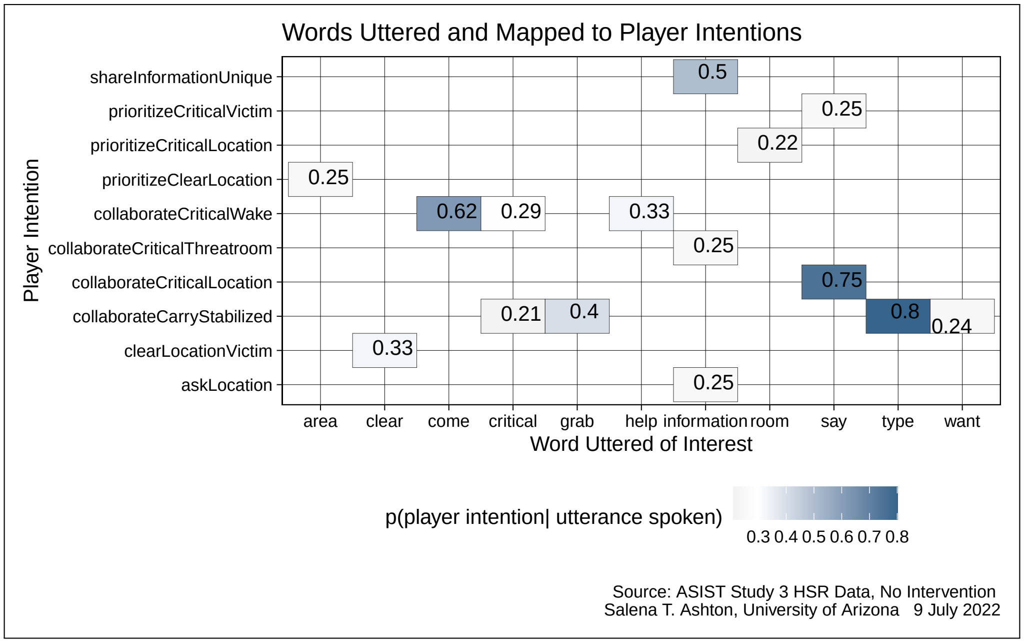
[12] It is not justified to claim 100% probability



Source: ASIST Study 3 HSR Data, No Intervention
Salena T. Ashton, University of Arizona 9 July 2022

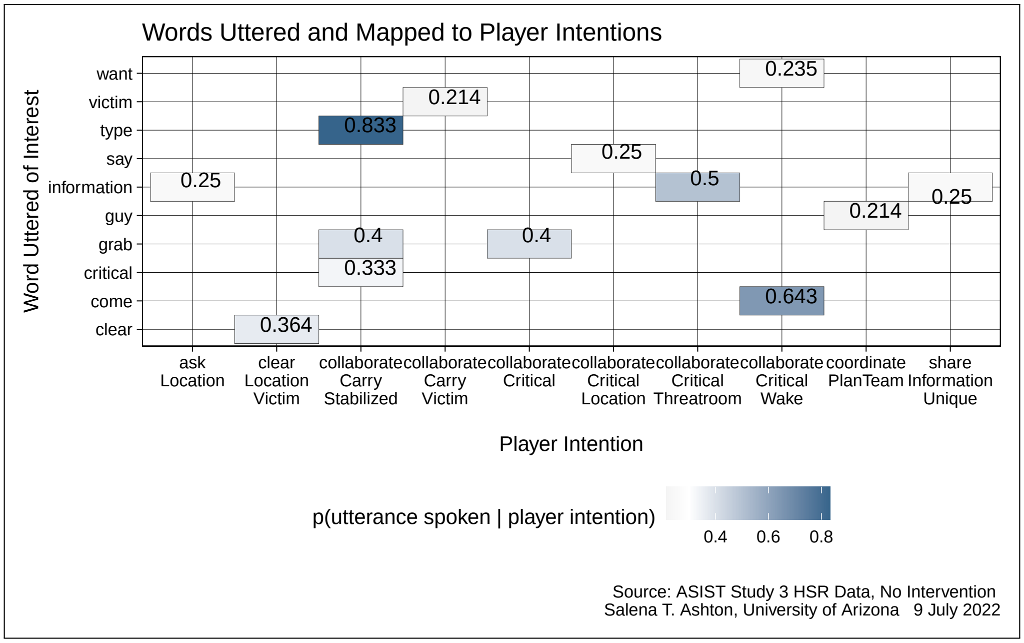
Figure 5: Questions asked with respect to time.

How to read this visual: Team 218 (videos 635 and 636) asked the fewest questions and scored the lowest combined point. this figure shows the lowest count of questions as a histogram. Video 636 is colored red in more than 66% of its bars. This means that of any questions the team did ask, more than 66% of them were asked toward the end of the mission. Compare this to team 1 (videos 633 and 634), who were relatively quiet players and asked the most questions in the beginning of each mission. Team 3 asked the most questions; as can be seen from the more even distribution of red and black, they asked questions at a consistent *rate* throughout their missions.



Source: ASIST Study 3 HSR Data, No Intervention
Salena T. Ashton, University of Arizona 9 July 2022

Figure 6


Words Uttered and Mapped to Player Intentions

Source: ASIST Study 3 HSR Data, No Intervention
Salena T. Ashton, University of Arizona 9 July 2022

Figure 7

5.5 Discussion and Future Research

The ultimate goal of this work is two-fold: find enough evidence to warrant the continued investigation of question-asking through causal reasoning to discern player intention and to algorithmitize player intention from player utterance transcriptions of player without human annotator dependency. Due to the small sample size, we do not offer conclusive evaluations at this time. However, from these 12 video annotations, we *can warrant* the further investigation of question-asking and inferred goals or intentions among human team players.

Key question phrases to give clues to ASI agents:

- ”If”
-

FUTURE RESEARCH:

- knowledge engineering and representation of tasks, goals, and hidden intentions; *hierarchical task network* (HTN),
- scale for additional data
- Preliminary research can lead to additional annotation, quantitative analysis and generalized theories for ASI
- Compare and contrast theories that emerge from no-intervention data to high-intervention data
- Investigation into which data structures best capture human individual and team ToM
- Promised Future Investigation of RESEARCH QUESTIONS:
  - How do spoken questions reveal another person’s plan or intent?
  - Can people listen to spoken questions and accurately decide if the other person’s plan is simple, sequential, or hierarchical? Can the distinction between such structures improve predictive performance for Artificial Social Intelligence (ASI) agent intervention?
- asdf