Towards Socially Intelligent Agents with Mental State Transition and Human Value

Liang Qiu[*1], **Yizhou Zhao**[*1], **Yuan Liang**[2], **Pan Lu**[1], **Weiyan Shi**[3], **Zhou Yu**[3], **Song-Chun Zhu**[1]

[1]UCLA Center for Vision, Cognition, Learning, and Autonomy
[2]University of California, Los Angeles
[3]Columbia University
`liangqiu@ucla.edu`

Abstract

Building a socially intelligent agent involves many challenges. One of which is to track the agent's mental state transition and teach the agent to make decisions guided by its value like a human. Towards this end, we propose to incorporate mental state simulation and value modeling into dialogue agents. First, we build a hybrid mental state parser that extracts information from both the dialogue and event observations and maintains a graphical representation of the agent's mind; Meanwhile, the transformer-based value model learns human preferences from the human value dataset, VALUENET. Empirical results show that the proposed model attains state-of-the-art performance on the dialogue/action/emotion prediction task in the fantasy text-adventure game dataset, LIGHT. We also show example cases to demonstrate: (*i*) how the proposed mental state parser can assist the agent's decision by grounding on the context like locations and objects, and (*ii*) how the value model can help the agent make decisions based on its personal priorities.

1 Introduction

Recently, there has been remarkable progress in language modeling with large-scale pretrained models (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019). Such models are used to build either general chatbots (Zhang et al., 2020) or task-oriented dialogue systems (Peng et al., 2020; Acharya et al., 2021; Qiu et al., 2020). While most of these systems have been able to generate fluent sentences, there are two major challenges towards building socially intelligent agents. First, considering dialogues as a "meeting of minds" (Gardenfors, 2014) or achieving some alignment of the interlocutors' mental models (Rumelhart et al., 1986; Stolk et al., 2016), few existing works are explicitly

Figure 1: Socially intelligent agents with mental state simulation and human values.

tracking the mental state transition of agents (Adhikari et al., 2020). Endowing current dialogue systems with such capability would allow the agent to condition its utterance on the context, simulate the effect of its actions, and further help understand the extended meaning, implicature, and irony expressed by the user (Grice, 1981, 1989). Second, it remains under-explored to teach agents to make a rational decision guided by its value. From a social and cultural perspective, humans tend to have a common preference described by the utility function related to individual values, common sense, and social awareness. For the example in Figure 1, someone who values personal security prefers staying at home rather than going outside at night.

Our work aims to alleviate the aforementioned problems, based on Embodied Cognitive Linguistics (ECL) (Lakoff and Johnson, 1980; Gardenfors, 2014) and established value theories in sociology (Schwartz, 2012). The ECL states that natural language is inherently executable, driven by mental simulation and metaphoric inference (Lakoff and Johnson, 1980), and learned through embodied interaction (Feldman and Narayanan, 2004; Tamari et al., 2020). Following its tenets, we present a hybrid mental state parser that converts dialogue and

---

[*]Equal contribution. The work was done prior to Liang joining Amazon Alexa.

event observations into a graphical representation of the agents' mind. Initialized with the location and object description, the interpretable representation is updated through the interaction history to track the evolving process of an agent's belief about surroundings and other agents.

In the field of intercultural research, Schwartz (1992); Schwartz et al. (2012) identify basic individual values that are recognized across cultures. Inspired by the theory, we propose to incorporate a value model that learns social common preferences from the human value knowledge base, VALUENET (Qiu et al., 2022). We perform experiments on a large-scale text-based embodied environment LIGHT (Urbanek et al., 2019). Empirical results show that the model with our mental state emulator and value function achieves the highest performance that aligns with human annotation among existing transformer-based models. Moreover, case studies further demonstrate that the mental state provides extra context information, while the value model helps agents make value-driven decisions.

Our contributions are two-fold. First, we propose to rethink the design of current dialogue systems and suggest a new paradigm from the perspective of cognitive science and contemporary sociology. Second, we present a new framework for building socially intelligent agents by incorporating mental state simulation and human value modeling into dialogue generation and decision making. Our methodology can be generalized to a wide range of interactive social situations in dialogue systems (Zhao, 2019), virtual reality (Lai et al., 2019), and human-robot interactions (Yuan and Li, 2017).

2 Related Work

2.1 Text-based Embodied AI

Most recent works in dialogues only study the statistical regularities of language data, without an explicit understanding of the underlying world. Virtual embodiment (Krishnaswamy and Pustejovsky, 2019) was proposed as a strategy for language research by several previous works (Brooks, 1991; Kiela et al., 2016; Gauthier and Mordatch, 2016; Mikolov et al., 2016; Lake et al., 2017). It implies that the best way to acquire human knowledge is to have the agent learn through experience in a situated environment. Urbanek et al. (2019) introduce LIGHT as a research platform for studying grounded dialogue (Grice, 1981, 1989; Stalnaker,

2002), where agents can perceive, emote, and act when conducting dialogues with other agents. Ammanabrolu et al. (2020) extend LIGHT with a dataset of "quests", aiming to create agents that both act and communicate with other agents in pursuit of a goal. Instead of guiding the agent to complete an in-game goal, our work aims to teach agents to speak/act in a socially intelligent way. Besides LIGHT, there are also other text-adventure game frameworks, such as Narasimhan et al. (2015) and TextWorld (Côté et al., 2018), but no human dialogues are incorporated in them. Based on the TextWorld, there are recent works (Yuan et al., 2018; Yin and May, 2019; Adolphs and Hofmann, 2019; Adhikari et al., 2020) on building agents trained with reinforcement learning.

2.2 Mental State Transition

An important hypothesis in the ECL (Lakoff and Johnson, 1980; Feldman and Narayanan, 2004) is that humans understand the meaning of language by mentally simulating its content. Great efforts have been made to model human mental states. For example, Dinan et al. (2019) design a memory network capable of storing knowledge and generating natural responses conditioning on retrieved entries. Adhikari et al. (2020) propose a graph-aided transformer agent (GATA) that infers and updates latent belief graphs during planning to enable effective action selection. However, GATA is designed for capturing game dynamics not dialogues, and our method is more flexible to encode both explicit environmental changes caused by agents' actions and implicit mental state updates triggered by agents' utterances. Such hybrid approaches mixing fixed symbolic states with deep continuous states are studied in recent neural-symbolic research (Sun, 1994; Garcez et al., 2008; Besold et al., 2017; Yi et al., 2018). The result interpretable graphs have two benefits: *(i)* the mental state parsing could be viewed as a form of executable semantic parses (Liang, 2016), so it is easy to write programs to simulate the mind transition. A real-world application leveraging similar approaches is seen in Andreas et al. (2020). *(ii)* the unified graphical representation can be extended to model higher-order mental states, *i.e.,* theory-of-mind (ToM) (Premack and Woodruff, 1978). ToM is defined as the ability to impute mental states to oneself and others. It enables humans to make inferences about what other people believe in a

given situation and predict what they will do (Apperly, 2010; Gordon and Hobbs, 2017; Akula et al., 2019). ToM is thus impossible without the capacity to form "second-order representations" (Dennett, 1978; Pylyshyn, 1978; Ganaie and Mudasir, 2015).

2.3 Human Value

When teaching agents to speak and act in a socially intelligent way, an approach considering values should be adopted. The theory of basic human values, developed by Schwartz (1992, 2012), tries to measure universal values that are recognized throughout major cultures. A set of 10 basic values[1] are identified and serve as the guiding principles in the life of a person or grou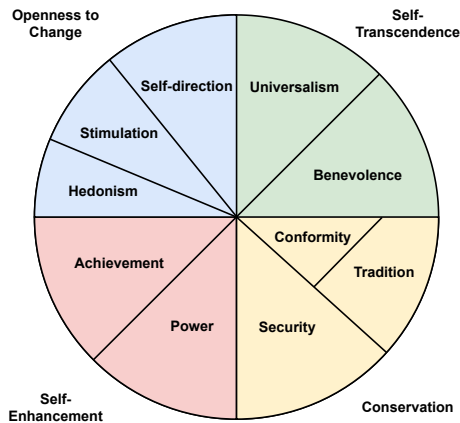p (Cieciuch and Davidov, 2012), as shown in Figure 2. Similarly, in economics and ethics, the concept of utility was developed as a measure of pleasure or satisfaction that drives human activities at all levels. Derived from the rational choice theory (Abella, 2009), utilitarianism states that human decision-making could be viewed as a two-step procedure. First, we select a feasible region based on financial, legal, physical, or emotional restrictions we are facing. Then we make a choice based on the preference order (Allingham, 2002; de Jonge, 2012). In this paper, we learn a transformer-based utility function of human values from the knowledge base VALUENET (Qiu et al., 2022). Inspired by descriptive ethics, VALUENET provides social scenarios and annotated human preference to teach the agent human attitudes to various ethical situations. The dataset is curated from the widely

Figure 2: Theory of Basic Human Values (Schwartz, 1992).

used social commonsense dataset SOCIAL-CHEM-101 (Forbes et al., 2020) and labeled with Amazon Mechanical Turk.

3 Problem Formulation

We will first briefly introduce the text-adventure environment LIGHT, followed by the mental state modeling and value utility formulation.

LIGHT (Urbanek et al., 2019) is a large-scale crowd-sourced fantasy text-adventure platform for studying grounded dialogues. Figure 4ⓐ shows a typical local environment setting, including location description, objects (and their affordances), characters, and their personas. Agents can talk to other agents in free-form text, take actions defined by templates, or express certain emotions (Figure 4ⓑ). Given the environmental setting and observation history, our task is to predict the agent's utterance/action/emotion for the next turn. To achieve this goal in a socially intelligent manner, we model the agent's mental state transition and incorporate human values. The mind model is proposed to depict the agent's belief about the underlying states of the text world. Meanwhile, a utility function of human values is designed to describe human preferences in common social situations. We experiment on the text-adventure game for simplicity, but the proposed architecture supports richer environments.
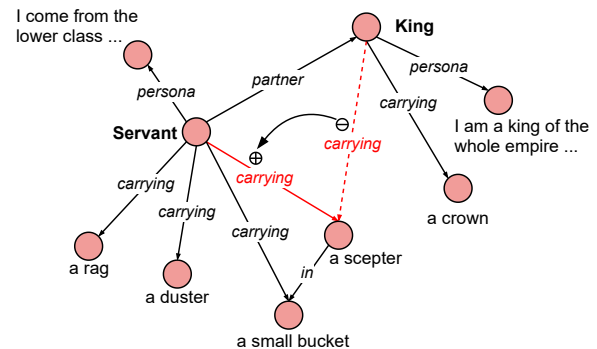

Figure 3: A graphical representation of the agent's mental state. Nodes are attributed with encoded natural language description of agents, objects and the environment. Agents' action trigger explicit topology changes of the graph.

3.1 Mental State Modeling

Our goal is to parse, construct and maintain the mental states in dialogues. With the mental state grounding on the details of the local environment, the agent could simulate and reason the evolution-

---

[1] A refinement of the theory (Schwartz et al., 2012), partitions the same continuum into 19 more narrowly defined values that permit more precise explanation and prediction.
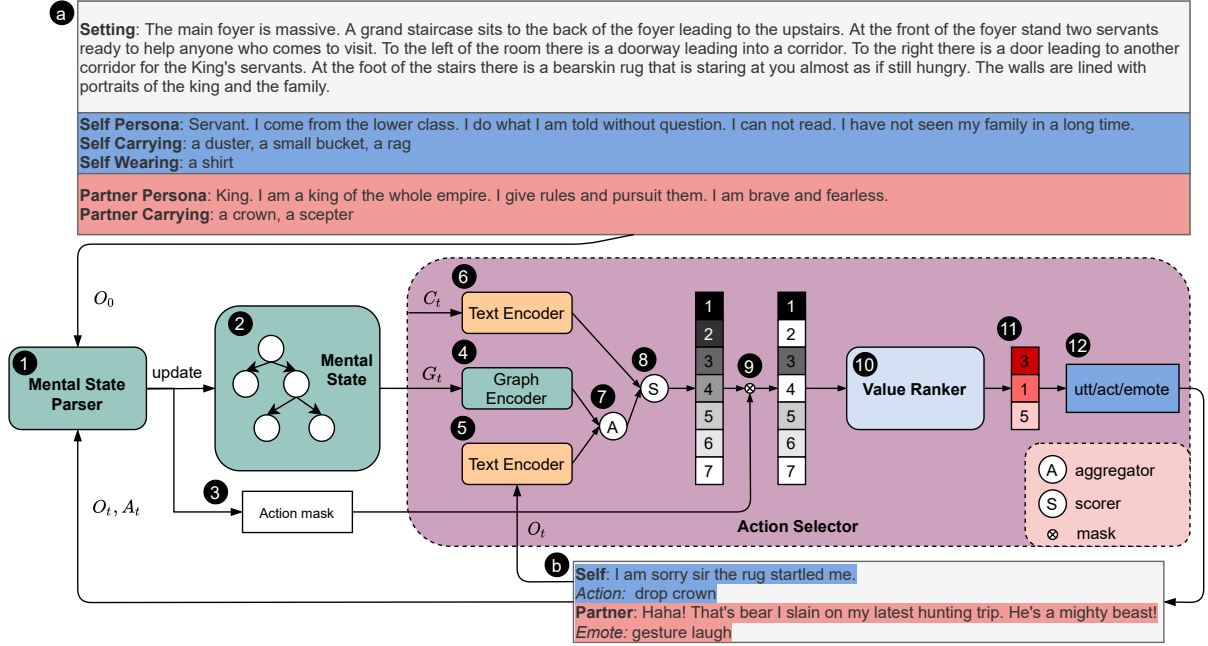
Figure 4: Socially Intelligent Agent Architecture with Mental State Parser and Value Model.

ary status of the world and condition its speaking and actions. A graphical representation of the mental state is proposed, as illustrated in Figure 3. Nodes in the graph represent the involved agents, persona descriptions, objects, objects' descriptions, and setting descriptions, which will change as the game setting switches. The relational edges between these nodes describe the state of mind. The mental state is updated with the observed dialogue history or actions, *e.g.*, *King gives the scepter to the servant* will result in the scepter being moved from the king to the servant.

3.2 Human Value Modeling

We assume that the agent in the fantasy world would make near-optimal choices to maximize the utility of its preferred values. We denote the available alternatives to be a set of n exhaustive and exclusive utterances or actions $A = \{a_1, ..., a_i, ..., a_n\}$. The value function $f_v(\cdot)$ describes the utility score of the alternative from the value dimension v, $v \in V = \{$*achievement, power, security, conformity, tradition, benevolence, universalism, self-direction, stimulation, hedonism*$\}$[2]. For example, if a_i is more preferred than a_j in terms of *security*, then $f_{security}(a_i) > f_{security}(a_j)$. Usually, we cannot find an analytical form of the value function. However, what matters for preference ordering is which of the two options gives the higher

expected utility, not the numerical values of those expected utilities.

In LIGHT, the agent's value priority is reflected by its persona description. For the example in Figure 4(a), the servant is a person who values *conformity* and *tradition* and has a lower priority on *self-direction* and *stimulation*. Using the same value function to approximate a value priority parser: $f_v(p)$, where p is the persona description, the utility or the desirability of candidate a_i to person p is the Euclidean distance between its value priority and the candidate's utility score:

$$u(a_i) = \sqrt{\sum_{v \in V} \left(f_v(p) - f_v(a_i) \right)^2}. \quad (1)$$

Since some actions could be impossible physically (*e.g., one cannot drop an object if the agent is not carrying the object*), the decision making process becomes a problem of maximizing the utility score that is subject to some constraints from the mental state, *i.e.*, $u(a|c)$, where c represents the context or constraints.

4 Algorithms

The overall architecture of our proposed framework is illustrated in Figure 4. For each scenario, a setting description (Figure 4(a)) is provided by the LIGHT environment, which can include a description of the location, object affordances, agents' personas, and the objects that agents are carrying,

---

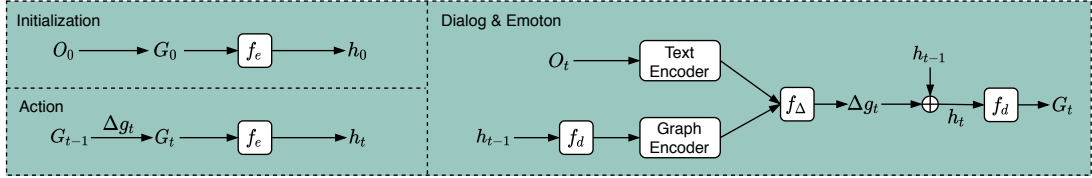[2]Detailed definition for each dimension is attached in Appendix A.1.

Figure 5: Overall Architecture of the Hybrid Mental State Parser

wearing, or wielding. The free-form conversations, actions, and emotions are logged during the communication as the observation history (Figure 4ⓑ). To begin with, a mental state parser will parse the setting descriptions into graph representation and initialize the agent's mental state (steps ① and ②). Besides the mental state updating, the parser also outputs an action mask that is aimed to rule out actions that are physically or causally impossible to take (step ③). A graph encoder (step ④) and a text encoder (step ⑤) will convert the mental state graph G_t and the dialogue observation O_t into vector representations, respectively. The same text encoder will be used to encode the candidates C_t (step ⑥). In step ⑦, the context vectors are combined by a bi-directional attention aggregator (Yu et al., 2018; Seo et al., 2016), and each candidate is assigned a score with a Multi-Layer Perceptron (MLP) (step ⑧). The action mask is then applied to get the feasible candidates under current mental state constraints (step ⑨). In steps ⑩ and ⑪, the top three candidates from the last step will be fed into the value model and re-ranked. Finally, the selected utterance/action/emotion is executed by the agent (step ⑫) and fed back to the environment. Upon receiving the response from other agents in the environment, the new observation will be again parsed and used to update the agent's state of mind, and the cycle repeats. In the following, we will describe each component in more detail.

4.1 Mental State Modeling (steps ①-②)

Figure 5 describes the architecture of the mental state parser. We define the mental state graph $G \in [-1, 1]^{R \times N \times N}$, where R is the maximum number of relation types and N is the maximum number of entities. The initial mental state graph G_0 is constructed by a ruled-based parser from the setting description O_0. The graph is encoded by function f_e to a hidden state h_0 that is later used for graph update. At game step t, the mental state parser parses relevant information from observation O_t and update the agent's mental state from G_{t-1} to G_t. Considering that observation O_t typ-

ically conveys incremental information from step $t-1$ to t, we generate the graph update Δg_t instead of the whole graph at each step

$$G_t = G_{t-1} \oplus \Delta g_t, \tag{2}$$

where $\oplus$ is the graph update operation. The graph update can be either discrete or continuous, and there have been studies on the pros and cons of each updating method (Adhikari et al., 2020). The discrete approach may suffer from an accumulation of errors but benefit from its interpretability. The continuous graph model needs to be trained from data, but it is more robust to possible errors. In this work, we propose a hybrid (discrete-continuous) method for updating the agent's state of mind by considering there exists a mixture of discrete events and continuous information in typical human-machine interactive environments. In the specific example of our tested LIGHT, the actions or events are template-based, it is more appropriate to adopt a discrete method for parsing; meanwhile, since utterances are challenging to be encoded into discrete representations, we apply a continuous update method instead.

4.1.1 Discrete Graph Definition & Update

To update the graph, we define Δg_t as a sequence of update operations of the following two atomic types:

- ADD(`src`, `dst`, `relation`): add a directed edge, named `relation`, from node `src` to node `dst`.

- DEL(`src`, `dst`, `relation`): delete a directed edge, named `relation`, from node `src` to node `dst`.

LIGHT defines various actions including *get*, *drop*, *put*, *give*, *steal*, *wear*, *remove*, *eat*, *drink*, *hug* and *hit*, and each taking either one or two arguments, *e.g.*, *give scepter to servant*. Every action could be parsed as one or a sequence of update operators that act on G_{t-1}. For example, actor performing "*give object to agent*" can be parsed into DEL(*actor*,

object, *carrying*) and ADD(*agent*, *object*, *carrying*). The rule-based parsing of the setting description and the discrete events could also be replaced by a seq2seq decoding process. Since both strings are well-structured in LIGHT, we omit training such a decoder for simplicity. Note that actions in LIGHT could only be executed when constraints are met, so we also generate an action mask according to the current mental state. By checking the adjacency matrix, we rule out action candidates conducted on objects that are inaccessible.

4.1.2 Continuous Graph Definition & Update

Besides the actions taken by the agents, their utterances could also have an implicit impact on the agents' mental states. To handle the continuous dialogue observation, we use a recurrent neural network as the graph update operation $\oplus$.

$$
\begin{aligned}
\Delta g_t &= f_\Delta(h_{G_{t-1}}, h_{O_t}), \\
h_t &= \text{RNN}(\Delta g_t, h_{t-1}), \quad (3) \\
G_t &= \text{MLP}(h_t).
\end{aligned}
$$

The function f_Δ aggregates the information from the previous mental state G_{t-1} and observation O_t to generate the graph update Δg_t. $h_{G_{t-1}}$ denotes the representation of G_{t-1} from the graph encoder. h_{O_t} is the output of the text encoder. h_t is a hidden state acting as the memory, from which we decode the new mental state G_t using a MLP. For the recurrent operator, we could either use LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014). More details on the graph encoder and text encoder we applied are presented in the section 4.2.

4.2 Action Selector (steps ④-⑪)

Conditioned on the agent's mental state, the action selector chooses the optimal candidate based on the prediction task (*i.e.*, utterance, action, or emotion). The selector consists of five components: a graph encoder (Fig. 4④) to convert the state-of-mind graph to a hidden state vector; a text encoder (Fig. 4(⑤, ⑥)) to encode the dialogue history and text candidates; an aggregator (Fig. 4⑦) to fuse the two context representations; a general scorer (Fig. 4⑧) to assign a score to each candidate; and a value model (Fig. 4⑩) to re-rank the candidates based on the assigned persona.

1. Graph Encoder. We use relational graph convolutional networks (R-GCNs) (Schlichtkrull et al., 2018) to encode the graph representation of mental states. The R-GCN is adapted from Graph Convolutional Networks (GCNs) so that it could embed the edge attributes (relational text embedding) in the mental state graph.

2. Text Encoder. A BERT-based (Devlin et al., 2019) encoder converts the text-based dialogue history into a vector representation, using the last hidden state corresponding to the `[CLS]` token; We also use the same encoder to encode the text response candidates.

3. Aggregator. A bi-directional attention layer (Yu et al., 2018; Seo et al., 2016) is adopted to fuse the information from the mental state and the contextualized text hidden state. The co-attention allows the agent to focus on the memory part that has been mentioned in the dialogue.

4. Scorer. The full context representation vector is concatenated with each candidate and an MLP layer with softmax activation generates a score for each of them.
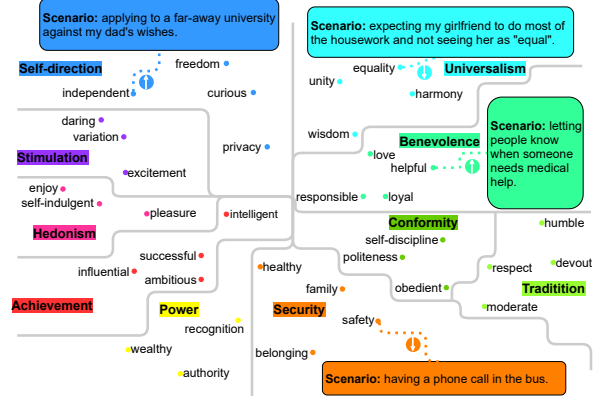

Figure 6: The VALUENET (Qiu et al., 2022) dataset with social scenarios organized by Schwartz values (Schwartz, 2012).

5. Value Ranker. After all the candidates are ranked, we select the top three candidates and then re-rank them according to the proposed value model. The value model is a BERT-based utility scorer trained on the knowledge base VALUENET (Qiu et al., 2022). A custom input format constructed as '`[CLS][$VALUE]s`' is fed into the BERT, *i.e.,*

$$
f_v(\texttt{s}) = \text{BERT}(\texttt{[CLS][\$VALUE]s}), \quad (4)
$$

where `[CLS]` is the special token for regression, `s` is the scenario, and `[$VALUE]` are special tokens we define to prompt (Li and Liang, 2021; Brown et al., 2020) the transformer the interested value dimension v. A regression head is put on top of the model to get a continuous estimation of the utility in the range of $[-1, 1]$.

Method	Seen Test			Unseen Test		
	Dialogue R@1/20	Action Acc	Emotion Acc	Dialogue R@1/20	Action Acc	Emotion Acc
BERT-based Bi-Ranker	76.5	42.5	25.0	70.5	38.8	25.7
BERT-based Cross-Ranker	74.9	50.7	25.8	69.7	51.8	28.6
discrete mental state	75.8	52.1	25.1	69.9	53.4	25.5
continuous mental state	77.3	49.3	**26.2**	72.1	45.2	29.1
hybrid mental state	78.4	53.5	26.1	72.3	54.3	29.5
hybrid+mask	78.5	54.5	26.1	72.3	55.4	29.4
hybrid+mask+value	**78.8**	**56.4**	26.1	**72.6**	**57.5**	**30.1**
Human Performance*	87.5	62.0	27.0	91.8	71.9	34.4

Table 1: Model performance on the LIGHT *Seen Test* and *Unseen Test*. For dialogue prediction, Recall@1/20 is reported for ranking the ground truth among 19 other randomly chosen candidates. Percentage accuracy is calculated for action and emotion prediction. (*) Human performance is reported by the original paper (Urbanek et al., 2019) on a subset of data.

The VALUENET is organized in 10 dimensions of Schwartz values, as shown in Figure 6. It consists of social scenarios curated from SOCIAL-CHEM-101 (Forbes et al., 2020). And the samples are annotated by Amazon Mechanical Turk workers, who are asked about their attitudes towards provided scenarios. For example, if you are someone who values *benevolence*, will you do or say: "today I buried and mourned a rat"? Their choices (yes, no, unrelated) are then quantified to numerical utilities: +1, -1, 0, respectively.

5 Experiments

We conduct experiments on the LIGHT dataset and compare our model with state-of-the-art methods based on two variants of BERT models. An ablation study is carried out to justify our model design, and a case study is performed to demonstrate how the proposed framework could help the agent ground upon the environment details and make value-driven decisions.

5.1 Experimental Setup and Implementation

The dialogues in LIGHT are split into *train* (8539), *valid* (500), *seen test* (1000), and *unseen test* (739) as the dataset is released. The *unseen test* set consists of dialogues collected on a set of scenarios that have not appeared in the training data. We use the history of dialogues, actions, and emotions to predict the agent's next turn. Note that the original paper manually filters out actions with no affordance leveraging the object annotation, while we provide all candidates to demonstrate our model's capability of reasoning feasible actions automatically from the agent's mental state.

Here we describe the implementation details of the proposed framework. The mental state graph is initialized with a structured setting string including all involved elements in the scenario (an example is attached in Appendix A.2). The setting parser is based on general parsing tools: regular expression and spaCy (Honnibal and Montani, 2017; Clark and Manning, 2016; Honnibal and Johnson, 2015), resulting in the initial mental state graph as shown in Figure 7. For the functions f_e and f_d, we
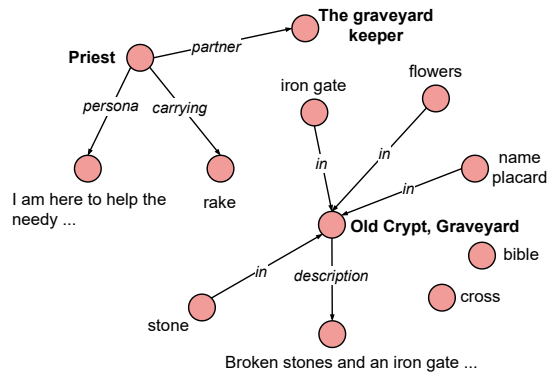

Figure 7: Initial mental state graph parsed from the example setting string in Appendix A.1. The nodes of objects' descriptions are omitted to save space.

use two-layer MLPs with tanh (Karlik and Olgac, 2011) and ReLU (Agarap, 2018) activations. The **Text Encoder** is a pretrained BERT (base-uncased) model (Wolf et al., 2020). The **Graph Encoder** is an R-GCN with six layers and a hidden size of 64. We also adopt the highway connections between consecutive layers for faster convergence and 3-basis decomposition to reduce the parameters and prevent overfitting.

5.2 Baseline Models

Two BERT-based models (Urbanek et al., 2019) are used as strong baselines, which have kept the state-of-the-art performance on this task. **BERT Bi-Ranker** produces a vector representation for the context and each candidate. Each candidate is assigned a score by the dot product between the context embedding and the candidate embedding. **BERT Cross-Ranker** concatenates the context string with each candidate and feeds the string to the BERT model instead. Compared with the bi-ranker, The cross-ranker allows the model to attend to the context when encoding each candidate.

5.3 Results and Analysis

Table 1 shows the results, where our model outperforms the state-of-the-art models by a large margin. To understand the results, we first compare mental state graph designs using discrete, continuous, and the proposed hybrid parser.

The discrete mental state parser uses actions to explicitly update the graph to augment the context representation. In the action prediction task, the discrete parser outperforms the purely continuous method (+2.8% (seen), +8.2% (unseen)), the BERT Bi-Ranker (+9.6% (seen), +14.6% (unseen)), and the BERT Cross-Ranker (+1.4% (seen), +1.6% (unseen)). While the continuous mental state parser misses the hard constraints introduced by less frequent actions, it updates the graph implicitly with the dialogues and shows a better result than the discrete one on dialogue prediction (+1.5% (seen), +2.2% (unseen)) and emotion prediction (+1.1% (seen), +3.6% (unseen)).

The hybrid mental state parser performs the best among the three according to almost all metrics, mainly because it aggregates the soft update from the dense dialogue and the hard constraints from the sparse actions. We also notice that the emotion prediction in LIGHT is a hard task because it is not strictly constrained by the context. Even humans can only achieve 27.0% (seen) and 34.4% (unseen) accuracy. Nevertheless, our model provides a relatively 1.2% (seen) and 3.1% (unseen) performance boost compared to the best BERT baseline.

Then, with the ablation study of our proposed action mask (hybrid mental state *vs.* hybrid+mask), we prove the effectiveness of it for improving action accuracy by ∼1% in action prediction. Figure 8 demonstrates how the mental state could help agent ground on the context. We can see a very
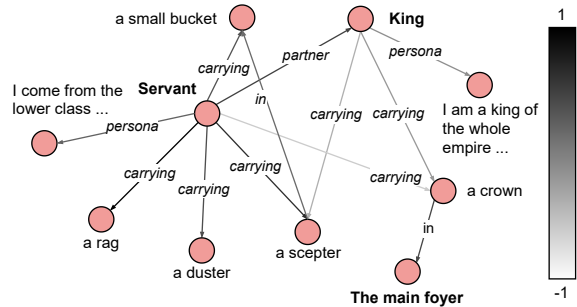

Figure 8: Intermediate mental state for the agent **Servant** in the dialogue example of Figure 4. The adjacency matrix of the mental state graph is visualized and the darkness of the edges represent the relation strength. Only critical relation types between nodes are shown for illustration purpose.

weak relation of the type "*carrying*" between the agent servant and the object crown. Thus the servant should not be able to give the crown to others at this time step. Though our model does not rely on annotated action affordances during action predicting, an action mask can be reasoned from such a mental state, which helps filter out physical or causally impossible actions.

Lastly, we analyze the results after introducing the value model. We first compute the value priority of the agent by applying the value function to its persona description. For example, given the servant's persona description in Figure 4, it shows *conformity, tradition*, and *security* have higher utility scores to the agent than other dimensions. Then we calculate utility scores of the top three candidates based on Equation 1. This teaches the agent to make decisions that align with the assigned role and further improves the overall performance, (+0.3% (seen), +0.3% (unseen)) for dialogue prediction, (+1.9% (seen), +2.1% (unseen)) for action prediction, and +0.7% (unseen) for emotion prediction.

6 Conclusion

This paper proposes to build a socially intelligent agent by incorporating mind simulation and human values. We explore using a hybrid parser to track agents' mental state transition. The value model pretrained on VALUENET brings social preference to help the agent make decisions. The model is proved to have a better performance than the state-of-the-art models on LIGHT. In the future, we have a plan to build a dataset to study the implicature in conversation and model deeper levels in the Theory of Mind based on the proposed representation.

References

Alex Abella. 2009. *Soldiers of reason: The RAND corporation and the rise of the American empire*. Houghton Mifflin Harcourt.

Anish Acharya, Suranjit Adhikari, Sanchit Agarwal, Vincent Auvray, Nehal Belgamwar, Arijit Biswas, Shubhra Chandra, Tagyoung Chung, Maryam Fazel-Zarandi, Raefer Gabriel, et al. 2021. Alexa conversations: An extensible data-driven approach for building task-oriented dialogue systems. *arXiv preprint arXiv:2104.09088*.

Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. 2020. Learning dynamic belief graphs to generalize on text-based games. *Advances in Neural Information Processing Systems*, 33.

Leonard Adolphs and Thomas Hofmann. 2019. Ledeepchef: Deep reinforcement learning agent for families of text-based games. *arXiv preprint arXiv:1909.01646*.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Arjun R Akula, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. 2019. X-tom: Explaining with theory-of-mind for gaining justified human trust. *arXiv preprint arXiv:1909.06907*.

Michael Allingham. 2002. *Choice theory: A very short introduction*. OUP Oxford.

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. 2020. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. *arXiv preprint arXiv:2010.00685*.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. 2017. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*.

Rodney A Brooks. 1991. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Jan Cieciuch and Eldad Davidov. 2012. A comparison of the invariance properties of the pvq-40 and the pvq-21 to measure human values across german and polish samples. In *Survey Research Methods*, volume 6, pages 37–48.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2018. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, pages 41–75. Springer.

Jan de Jonge. 2012. Rational and moral action. In *Rethinking Rational Choice Theory*, pages 199–206. Springer.

Daniel C Dennett. 1978. Beliefs about beliefs [p&w, sr&b]. *Behavioral and Brain sciences*, 1(4):568–570.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jerome Feldman and Srinivas Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and language*, 89(2):385–392.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

MY Ganaie and Hafiz Mudasir. 2015. A study of social intelligence & academic achievement of college students of district srinagar, j&k, india. *Journal of American Science*, 11(3):23–27.

Artur SD'Avila Garcez, Luis C Lamb, and Dov M Gabbay. 2008. *Neural-symbolic cognitive reasoning*. Springer Science & Business Media.

Peter Gardenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.

Jon Gauthier and Igor Mordatch. 2016. A paradigm for situated and goal-driven language learning. *arXiv preprint arXiv:1610.03585*.

Andrew S Gordon and Jerry R Hobbs. 2017. *A formal theory of commonsense psychology: How people think people think*. Cambridge University Press.

H Paul Grice. 1981. Presupposition and conversational implicature. *Radical pragmatics*, 183.

H Paul Grice. 1989. Indicative conditionals. *Studies in the Way of Words*, pages 58–85.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Bekir Karlik and A Vehbi Olgac. 2011. Performance analysis of various activation functions in generalized mlp architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4):111–122.

Douwe Kiela, Luana Bulat, Anita L Vero, and Stephen Clark. 2016. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *arXiv preprint arXiv:1610.07432*.

Nikhil Krishnaswamy and James Pustejovsky. 2019. Multimodal continuation-style architectures for human-robot interaction. *arXiv preprint arXiv:1909.08161*.

Hsu-Chao Lai, Hong-Han Shuai, De-Nian Yang, Jiun-Long Huang, Wang-Chien Lee, and Philip S Yu. 2019. Social-aware vr configuration recommendation via multi-feedback coupled tensor factorization. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1773–1782.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Percy Liang. 2016. Learning executable semantic parsers for natural language understanding. *Communications of the ACM*, 59(9):68–76.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 29–61. Springer.

Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Zenon W Pylyshyn. 1978. When is attribution of beliefs justified?[p&w]. *Behavioral and brain sciences*, 1(4):592–593.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191.

Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-Chun Zhu. 2020. Structured attention for unsupervised dialogue structure induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1889–1899, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

David E Rumelhart, Paul Smolensky, James L McClelland, and G Hinton. 1986. Sequential thought processes in pdp models. *Parallel distributed processing: explorations in the microstructures of cognition*, 2:3–57.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Arjen Stolk, Lennart Verhagen, and Ivan Toni. 2016. Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, 20(3):180–191.

Ron Sun. 1994. *Integrating rules and connectionism for robust commonsense reasoning*. John Wiley & Sons, Inc.

Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf. 2020. Language (re)modelling: Towards embodied language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6268–6281, Online. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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 673–683, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*.

Xusen Yin and Jonathan May. 2019. Comprehensible context-driven text game playing. In *2019 IEEE Conference on Games (CoG)*, pages 1–8. IEEE.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Wang Yuan and Zhijun Li. 2017. Development of a human-friendly robot for socially aware human-robot interaction. In *2017 2nd International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 76–81. IEEE.

Xingdi Yuan, Marc-Alexandre Côté, Alessandro Sordoni, Romain Laroche, Remi Tachet des Combes, Matthew Hausknecht, and Adam Trischler. 2018. Counting to explore and generalize in text-based games. *arXiv preprint arXiv:1806.11525*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Ran Zhao. 2019. *Socially-Aware Dialogue System*. Ph.D. thesis, Carnegie Mellon University.

A Appendix

A.1 Schwartz Value Definition

Self-Direction Defining goal: independent thought and action–choosing, creating, exploring. Self-direction derives from organismic needs for control and mastery and interactional requirements of autonomy and independence. (creativity, freedom, choosing own goals, curious, independent) [self-respect, intelligent, privacy]

Stimulation Defining goal: excitement, novelty, and challenge in life. Stimulation values derive from the organismic need for variety and stimulation in order to maintain an optimal, positive, rather than threatening, level of activation. This need probably relates to the needs underlying self-direction values. (a varied life, an exciting life, daring)

Hedonism Defining goal: pleasure or sensuous gratification for oneself. Hedonism values derive from organismic needs and the pleasure associated with satisfying them. Theorists from many disciplines mention hedonism. (pleasure, enjoying life, self-indulgent)

Achievement Defining goal: personal success through demonstrating competence according to social standards. Competent performance that generates resources is necessary for individuals to survive and for groups and institutions to reach their objectives. As defined here, achievement values emphasize demonstrating competence in terms of prevailing cultural standards, thereby obtaining social approval. (ambitious, successful, capable, influential) [intelligent, self-respect, social recognition]

Power Defining goal: social status and prestige, control or dominance over people and resources. The functioning of social institutions apparently requires some degree of status differentiation. A dominance/submission dimension emerges in most empirical analyses of interpersonal relations both within and across cultures. To justify this fact of social life and to motivate group members to accept it, groups must treat power as a value. Power values may also be transformations of individual needs for dominance and control. Value analysts have mentioned power values as well. (authority, wealth, social power) [preserving my public image, social recognition]

Both power and achievement values focus on social esteem. However, achievement values (e.g., ambitious) emphasize the active demonstration of successful performance in concrete interaction, whereas power values (e.g., authority, wealth) emphasize the attainment or preservation of a dominant position within the more general social system.

Security Defining goal: safety, harmony, and stability of society, of relationships, and of self. Security values derive from basic individual and group requirements. Some security values serve primarily individual interests (e.g., clean), others wider group interests (e.g., national security). Even the latter, however, express, to a significant degree, the goal of security for self or those with whom one identifies. (social order, family security, national security, clean, reciprocation of favors) [healthy, moderate, sense of belonging]

Conformity Defining goal: restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms. Conformity values derive from the requirement that individuals inhibit inclinations that might disrupt and undermine smooth interaction and group functioning. As I define them, conformity values emphasize self-restraint in everyday interaction, usually with close others. (obedient, self-discipline, politeness, honoring parents and elders) [loyal, responsible]

Tradition Defining goal: respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides. Groups everywhere develop practices, symbols, ideas, and beliefs that represent their shared experience and fate. These become sanctioned as valued group customs and traditions. They symbolize the group's solidarity, express its unique worth, and contribute to its survival (Durkheim, 1912/1954; Parsons, 1951). They often take the form of religious rites, beliefs, and norms of behavior. (respect for tradition, humble, devout, accepting my portion in life) [moderate, spiritual life]

Tradition and conformity values are especially close motivationally; they share the goal of subordinating the self to socially imposed expectations. They differ primarily in the objects to which one subordinates the self. Conformity entails subordination to persons with whom one frequently interacts—parents, teachers, and bosses. Tradition entails subordination to more abstract objects—religious and cultural customs and ideas. As a corollary, conformity values exhort responsiveness to current, possibly changing expecta-

tions. Tradition values demand responsiveness to immutable expectations from the past.

Benevolence Defining goal: preserving and enhancing the welfare of those with whom one is in frequent personal contact (the 'in-group'). Benevolence values derive from the basic requirement for smooth group functioning and from the organismic need for affiliation. Most critical are relations within the family and other primary groups. Benevolence values emphasize voluntary concern for others' welfare. (helpful, honest, forgiving, responsible, loyal, true friendship, mature love) [sense of belonging, meaning in life, a spiritual life].

Benevolence and conformity values both promote cooperative and supportive social relations. However, benevolence values provide an internalized motivational base for such behavior. In contrast, conformity values promote cooperation in order to avoid negative outcomes for self. Both values may motivate the same helpful act, separately or together.

Universalism Defining goal: understanding, appreciation, tolerance, and protection for the welfare of all people and for nature. This contrasts with the in-group focus of benevolence values. Universalism values derive from survival needs of individuals and groups. But people do not recognize these needs until they encounter others beyond the extended primary group and until they become aware of the scarcity of natural resources. People may then realize that failure to accept others who are different and treat them justly will lead to life-threatening strife. They may also realize that failure to protect the natural environment will lead to the destruction of the resources on which life depends. Universalism combines two subtypes of concern—for the welfare of those in the larger society and world and for nature (broadminded, social justice, equality, world at peace, world of beauty, unity with nature, wisdom, protecting the environment)[inner harmony, a spiritual life]

A.2 Example Environment Setting

An example setting string for the utterance prediction is:
"**_task_speech**
_setting_name Old Crypt, Graveyard
_setting_desc Broken stones and a iron gate closing the entrance with a name placard that the name is worn off.
_partner_name the graveyard keeper who lives across the yard **_self_name** priest
_self_persona I am here to help the needy. I am well respected in the town. I can not accept lying.
_object_desc a gate : The gate is made out of rusty metal. It squeaks as it swings on its hinges.
_object_desc a flowers : you can see them up close but not afar. when noticed, you realize that they are old.
_object_desc a name placard : The placard is made of wood witha clear name on it.
_object_desc a stone : The stone is chipped from being used as target practice from soldier trainees
_object_desc a placard : A sign used to display names of buildings or notices.
_object_desc an iron gate : The gate is ornate, with complicated iron scrollwork patterns.
_object_desc a Rake : This rake is made of carefully split wood with a sturdy looking handle. Seems useful for keeping the leaves under control.
_object_desc a Cross : The cross is broken and with a few dents in the sides.
_object_desc a bible : The bible is bound by black leather, its pages yellowed by years of use.
_object_in_room a gate
_object_in_room a flowers
_object_in_room a name placard
_object_in_room a stone
_object_in_room a placard
_object_in_room an iron gate
_object_carrying a Rake".
The result mental state graph parsed from this setting is illustrated in Figure 7.