Title

A Relevance-based Decision-making Model of Human Sparse, Overloaded, and Indirect Communication

Author

Jiang, Kaiwen

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

A Relevance-based Decision-making Model of Human Sparse, Overloaded, and Indirect Communication

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Kaiwen Jiang

2024

ABSTRACT OF THE DISSERTATION

A Relevance-based Decision-making Model
of Human Sparse, Overloaded, and Indirect Communication

by

Kaiwen Jiang
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2024
Professor Tao Gao, Chair

Human real-time communication creates a limitation on the flow of information, which requires the transfer of carefully chosen and concise data in various situations. Although pointing is sparse, overloaded, and indirect, it allows humans to effectively decode shared information, (ex)change their minds, and plan accordingly. I introduce a model that explains how humans choose information for communication and understand communication by utilizing the linguistics concept of "relevance" derived from decision-making theory and theory of mind.

The modeling approach taken in this dissertation is inspired by many seemingly separated domains. First, I apply theory of mind from cognitive science and partially observable Markov decision process to formally model the components of human mind and how they make decisions, building a scaffold for modeling human communication. Second, I derive how humans coordinate and share their mind by applying the concepts of paternalistic helping in developmental psychology and philosophical discussion about empathy. Third, I derived the definition of utility-based relevance as how much a signaler's belief can make a positive

difference to its receiver's well-being, utilizing the cooperative assumption of human communication in linguistics and comparative psychology. I conducted simulation and human behavioral experiments to show that relevance-based communication model can model the overloaded and indirect human communication and can predict humans' choices of signals in communication. Artificial intelligence agents that communicate with relevance-based models are more well-received by humans. Finally, I use Markov decision process and partially observable Markov decision process to propose a way of finding the best timing for sparse human communication.

The dissertation of Kaiwen Jiang is approved.

Richard Alan Clarke Dale

Hongjing Lu

Yingnian Wu

Tao Gao, Committee Chair

University of California, Los Angeles

2024

To all of my friends

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would never imagine myself completing this Ph.D. journey without the support and guidance from many incredibly amazing people. I want to say thank you.

I want to thank my advisor, Dr. Tao Gao, for guiding me in research, career, and life. You lead me not only to my love in research, modeling human behavior, but also to my passion on education. You showed me what a passionate scientist is and how to devote to a passion. You also led me through the difficult times when I was at a loss in research. I will miss your small talks on random topics like American history and renaissance arts.

I want to thank my friends in the visual intelligence lab. I want to thank Stephanie Stacy for setting a role model for me. I cannot express how happy and surprised I am to see you at my defense. I want to thank Minglu Zhao for always being a great friend and always having fun chats in and out of research, Siyi Gong for all the moments we shared in discussions and dinners, Aishni Parab for always being caring and supportive, Victor Zhang for always leading me to exploring new ideas and developing new skills. I want to thank my undergraduate collaborators as well: Boxuan Jiang, Annya Dahmani, Adelpha Chan, Chuyu Wei, Anahita Sadaghdar, and Rebekah Limb. I enjoyed working with you and learned so much from you.

I want to thank my mentor friends. My mentor in General Electric Research, Peter Tu, is so insightful in both research and life philosophy. We had so many quality times together. I also appreciate the Professors that I worked together as a teaching assistant and staffs in the Department of Statistics and Data Science. Michael Tsiang, Linda Zanontian, Guani Wu, Rick Schoenberg, Rob Gould, and Chie Ryu, thank you for helping me learn teaching, providing suggestions when I teach my course, and all the helps with my career.

I want to thank my friends who supported me all along the journey. Dehong Xu and Jiaming Guo, we supported each other through the hard Covid times. Yaqi Han, Simon Wu, and Xu Han, you gave me so much company and laughter. Chuang Lan, Shensi Xiao,

and Tianxiong Ju, you have been wonderful friends ever since childhood. I cherish all the moments with you.

Finally, I want to thank my family for being unconditional supportive. Despite the distance, our hearts are always connected.

Portions of this work are adapted from collaborative works. Contributions to this manuscript are as follows. Parts of Chapter 1 and 2 are adapted from a. SS drafted parts of cognitive modeling. SG drafted parts focusing on developmental and comparative psychology and made editing contributions. MZ and AP contributed to edits on modeling cooperation, and TG made theoretical contributions. Chapter 3 is expanded from Jiang et al. (2021) AC and CW helped drafting the discussion in philosophy and developmental psychology and editing through the manuscript. TG made theoretical contributions. FR and YZ gave valuable feedback. Chapter 4 is expanded from Jiang et al. (2022) made many theoretical contributions to the model, AD and BJ contributed to drafting of the parts in cognitive and developmental psychology. Chapter 5 is adapted from a manuscript with Boxuan Jiang, Anahita Sadaghdar, Rebekah Limb, and Tao Gao. BJ helped designing and conducting the experiments, TG provided suggestions to developing the experiment, AS and RL contributed to conducting the experiments and editing the manuscript. Chapter 6 is a work together with Thomas Li and Tao Gao. TG provided theoretical suggestions. TL helped design maps and heuristic models and conducting simulation experiments.

VITA

2011–2015 B.S. Psychology and Statistics, Peking University

2016–2018 M.S. Applied Statistics, University of Michigan, Ann Arbor

2018–2019 Biostatistician Specialist, Mylan Pharmaceuticals

2019–2024 Teaching Fellow, UCLA Department of Statistics and Data Science

2022 Computer Vision Intern, General Electric Research

PUBLICATIONS

Jiang, K., Stacy, S., Chan, A., Wei, C., Rossano, F., Zhu, Y., & Gao, T. (2021). Individual vs. Joint Perception: a Pragmatic Model of Pointing as Smithian Helping. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 43, No. 43).

Jiang, K., Stacy, S., Dahmani, A. L., Jiang, B., Rossano, F., Zhu, Y., & Gao, T. (2022). What is the point? a theory of mind model of relevance. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 44, No. 44).

Stacy, S., Gong, S., Parab, A., Zhao, M., Jiang, K., & Gao, T. (2024). A Bayesian theory of mind approach to modeling cooperation and communication. Wiley Interdisciplinary Reviews: Computational Statistics, 16(1), e1631.

CHAPTER 1

Introduction

Humans have fascinating ability to communicate efficiently, a fact that we tend to ignore. Let's look at an example. Tom works at the information desk in a library. Emma is waiting in the front of the line when Tom waves at her. Emma goes to Tom and asks: Where can I find *Gone with the Wind*? Tom puts a finger over his lips, points to the computer and starts to look up in the system. Emma keeps silent while Tom sends a text "Shelf H, row 25" to his co-worker Cindy. Tom points to Emma when Cindy retrieves the book. Emma gives Tom her membership card and checks the book out. On her way out of the library, Emma sends to Chat-GPT: Can you give me some background information on *Gone with the Wind*?

Conversations, texts, gestures, pointing, eye gazing, and even silence, multiple means of communication are used between individuals, and growingly, between individuals and artificial intelligent systems. Through linguistic or visual, ostensive or obscure communication, massive information is obtained and exchanged by human through in a swift and efficient way. With communication, we can intuitively learn what are on people's mind: their goal, plan, knowledge, and what they are paying attention to; We can establish a mutuality that we are paying attention to the same matter and some information are about to be made public; We can gather a massive amount of data to train artificial intelligent systems to in turn communicate with humans in a more human compatible way.

It is natural to ask, what makes people communicate? How can such impromptu and momentary communication signals efficiently convey their meaning? What role is communication playing in people's lives? To address these questions, I propose a relevance-based

computational model of human communication with insights from cognitive science, developmental psychology, and reinforcement learning. We start by investigating the characteristics of human communication.

1.1 Characteristics of Human Communication

1.1.1 Sparsity of Human Communication

Human communication conveys massive information in an incredibly disproportionally short amount of time. In a volleyball game, the setter gestures behind their back to tell the strategy for this point, indicating the movement plan for the next couple of movements; In a surgery, the surgeon raise his hand to the nurse to demand a specific tool among a collection of more than 10 tools; In a dinner date, the latecomer sends a text 5 minutes to inform that they will arrive in 5 minutes. All these communications happen within a mere second and with a low frequency. We say that human communication is **sparse**.

In daily life, we seldom entirely say everything we mean to say. It is hard to imagine the surgeon saying "I want the clamp on the left side. Can you hand it over and put it in my left hand?" in the surgery, or the latecomer typing "I got in a traffic. I will be at the Nando's restaurant at Plymouth and Broadway with you in 5 minutes from now. Please be patient." Actually, if we hear people speak like this, we will question the expertise or development of the listener. A convention is that common knowledge should not communicated as we all know it and we know the fact that we all know it (Campbell, 2005). The surgeon saying that they want the clamp indicates that the nurse may not know what the clamp is or that the clamp is needed at this step of the procedure; The latecomer texting the name of the restaurant indicates that the partner may not know which restaurant they reserved. We only mention this common knowledge to inform people who has not access the knowledge, for example, teaching students, or explaining insider jokes.

The sparsity of human communication also reflects in the lack of repetitiveness of human

communication. While humans involve in interpersonal communication very frequently, we do not tend to repeat communicating about the same matter frequently to the same listeners. If we have communicated and I know that you have understood, the message is put into common knowledge. Therefore, we should not communicate about it again. We only repeat messages if we do not expect the message has been received and understood, or we want to mark the message to emphasize. For example, after a father tells his kid that broccoli is good for health, he will only say it again if he sees the kid picking out broccoli from lunch, which indicates that the kid may still not take the nutrition value of broccoli seriously.

1.1.2 Overloaded Human Communication

With such sparse and condensed signals, human communication usually has multiple possible referent and meanings. We take a form of visual communication, pointing, as an example. As Wittgenstein and Anscombe (1953/2001) stated: "Point to a piece of paper. Now to its color, to its shape." You will find that these three pointing signals are the same: If you point to a piece of paper, the signal is merely an extension of the finger. Perceived by the receiver, the pointing may refer to a set of objects in the environment that lie on this extension. Maybe you are pointing to the piece of paper. Maybe you are pointing to the table on which the paper sits. Maybe you are pointing to the pen that is next to the paper. Adding to the overloaded nature of pointing, multiple meanings are possible even if we identify the referent of pointing. Given that the pointing refers to the paper, it is not clear whether the pointing is to its color, its shape, or the fact that it was written on. We can call this one-to-many nature of mapping from signal to referent/meaning to be **overloaded**.

The overloaded nature of pointing signal does not only apply to pointing, but both visual and verbal communication. In the examples in Chapter 1.1.1, the signals are all overloaded, even when they have a conventionalized literal meaning. The gestures with one finger, two or three fingers sticking out may mean number 1, 2, or 3, but on the volleyball court they are decoded as various strategies. Raising a hand may mean *I want a tool*, or *I don't*

have anything, or even *my arm aches*, but in the surgeon room it means *give me a specific tool*. The 5 minutes text may have so different meanings: *I will be there in 5 minutes*, *the party will start in 5 minutes*, or *the bus will come in 5 minutes*. Technically, we use every communicative signal with multiple different meanings. Besides the literal meaning of a sentence or gesture, different contexts, tones, repetition, and markedness also influence how people interpret the signals, making sarcasm or humor possible. Misyak, Noguchi, and Chater (2016) demonstrates that human pointing is extremely impromptu and overloaded. Based on slightly different game settings, the same pointing gesture can have multiple different meanings. These meanings can communicate beyond the object's visual features to what to do with the referent and why the pointer used such a gesture.

To be even more extreme, some communicative signals do not have a literal meaning. For example, when someone at the street suddenly shout out *ahh*. Based on this signal, you may think there might be something wrong. It could be a gutter on the road or an instruction site blocking the road. This single ahh sound may have two completely opposite meanings, but humans can effectively use the signal to communicate briskly in real life. We will further discuss overloaded communicative signals in Chapter 4.

1.1.3 Indirectness of Human Communication

Human communication is also indirect: Talking about something may have a completely different meaning than doing something directly with the referent. Instead, it can go far beyond the referent visual feature or object. When a lifeguard at the beach points to a cloud to you, he may mean that a rain is approaching, and you should leave. Neither the protagonist or the lifeguard can do anything to the cloud. The lifeguard's meaning has no direct relationship with the cloud. In a negotiation, a lead representative says *I would like to use the restroom*. It may have nothing to do with the restroom, but both parties know that the representative wants to leave the room for a moment to discuss with the group members.

The indirectness of human communicative signals exploits the overloaded nature of com-

munication. Direct or indirect, a signal can have infinite many possible meanings. However fortunately, humans can efficiently employ and interpret the meanings of these sparse, overloaded, and indirect signals with an incredibly high accuracy and flexibility (Kendon, 2004; Lascarides & Stone, 2009). The overloaded signals may have different meanings. However, once the context or scenario where the signal is employed is fixed, their meaning is fixed as well. Instead of a many-to-one mapping from signal to meaning, there appears to be a many-to-one mapping from (signal, context) to meaning. In the next section, we review models of communication and investigate how context is studied in the discussion between communicative signals and meanings.

1.2 Models of Communication

1.2.1 Code-based Perspective: a Fixed Codebook

Communication is long seen as a code transmission from a sender to a receiver, based on Shannon's classical communication model (Shannon, 1948). A code w is encoded by the sender from a meaning m, passes through a noisy channel, gets received by the receiver as $\hat{w}$, and decoded by the receiver as a received meaning $\hat{m}$ with a decoder. Even though some models consider the probabilistic encoder $p(w|m)$ and decoder $q(\hat{m}|\hat{w})$, most Shannon communicative models have a deterministic codebook defines a mapping between signals and meanings. In these models, the uncertainty in the transmission comes from the degradation of the signals when passing through the noisy communicative channel. The focus of these models is to design a codebook that a) enables an accurate retrieval of signal from a noisy channel or b) has the shortest code in expected length (MacKay, 2003). These models do not concern the uncertainty introduced by the relationship between meanings and signals, therefore do not model overloaded human communication to a large extent.

Many models in AI adopt this perspective and treat communicative signals as codes. For example, in many communicative scenarios in partially observable environments, communi-

cation is to directly share a copy of the observation from the sensors of each individual agents (Spaan, Gordon, & Vlassis, 2006; Xuan, Lesser, & Zilberstein, 2001). In research on AI assistants, the agents also communicate with observations in the environment (Reddy, Levine, & Dragan, 2021). These models focus designing AI agents that can efficiently interact with each other, or between the AI agent and a human individual, so the communication is not necessarily human-like. Even though in some of these models, the communicative signals employed by the AI agents even show language properties (Havrylov & Titov, 2017), they fail to flexibly capture the overloaded nature of human communication, like pointing, eye gazing, or a shout.

1.2.2 Context-dependent Overloaded Communication

In this section, we review communication models which discuss the relationship between signals (utterances), meanings, and context. We aim to see how context is defined in each of the models.

1.2.2.1 Computational linguistics: what has been said

The interpretation of human communicative signals is not only dependent on the signal itself, but also depending on what has been said before and what is said after it. For example, the word bank has two different meanings. If in the texts before and after the word bank, more words like money, federal, or stream are used, then it is more likely to be interpreted as a financial establishment; if in the context, words like stream, river, or deep are used, then the word is more likely to have the meaning as the land alongside a river (Griffiths, Steyvers, & Tenenbaum, 2007).

In most computational linguistics models, the context is defined as the words before (and after) the current word or sentence. It could be a window around the current word or the sentence where the current word locates (Collobert et al., 2011). For example, in

probabilistic grammar, a context is modeled as the words $w_{1:i-1}$ before the current world w_i and the condition probability $P(w_i|w_{1:i-1})$ is studied to predict the next word. In natural language processing modeled like n-grams, the context is the $n-1$ words before the current word (Damashek, 1995). In state-of-the-art language models transformers (Vaswani et al., 2017) and further contextual word embedding models like ELMo (Peters et al., 2018) and BERT (Devlin, Chang, Lee, & Toutanova, 2018), the context is converted to embedding vectors and weights are assigned to each word before and after the current word or sentence. With these weights, the attention mechanism can be used to predict the next word or the masked word.

Some language models achieve a fascinating level in understanding overloaded human language and hence building question-answering AIs (OpenAI, 2023). However, these models do not address the sparsity of human communication in two aspects. First, when the communicative signal does not have a long sequence of words ahead, for example, upon receiving the text *on my way*, it is hard for these models to predict the referent of this text. Also, the overemphasis of texts before and after the current utterance in these models results in overly lengthy answers provided by these models.

1.2.2.2 Gricean communication: what could have been said

Studying human communication, Grice (1975) proposed an influential set of principles on linguistic pragmatics called the Gricean maxims. According to the Gricean maxims, communication should be truthful, concise, relevant, and straightforward, under the umbrella term cooperative principle. The framework proposes that a signal is used to convey information about the states of the world in a maximally efficient way. The listener does not only understand what the speaker literally means by the words, but also try to infer an implicated meaning under the assumption that the speaker follows the maxims. These are the literal meaning (or semantic meaning) and pragmatic meaning of an utterance. The pragmatic meaning is uncovered by the listener considering what has been said and what could have

been said, which is the context.

On top of the Gricean communication maxims, Frank and Goodman (2012) proposed the rational speech act (RSA) model. RSA models human communication as rational actions of the speaker and perceived observations of the listener. It originates from a language game called reference games in which the speaker selects an utterance from a set of utterances and sends to the receiver to help identify one target from a set of distractors. In RSA, the context is what has been said – the utterance and what could have been said – the other utterances in the utterance set. For example, in a set of objects green square, green circle, blue circle, if the speaker sends out the utterance *green*, the listener is more likely to think the target is the green circle than it is the green square. The key to reasoning is: if the target were the green square, the speaker could have said *square* so that the ambiguity is eliminated. In this example, what has been said, *green*, and what could have been said, *square*, together determine the listener's interpretation of the utterance. We will introduce the probabilistic model of RSA in detail in Chapter 2.

Gricean communication and RSA model context beyond simply what has been said in the computational linguistic models, but also include what could have been said. This expands the interpreted meanings of the utterances from literal meanings to pragmatic meanings. However, in these models, the receiver's interpretation of the utterance highly depends on comparing the literal meaning of the utterances. For example, in reference games with RSA, the utterance *green* eliminates the object blue circle from the target because the utterance is not consistent with the features of the blue circle. However, beyond language games, these models would have a hard time interpreting utterances whose literal meaning is not stark apparent, like pointing, eye gazing, or just a simple *Ahh*.

1.2.2.3 Cooperative motivated communication: why it is said

The third perspective studies the development of human communication from its origin: the human-unique cooperation. In this perspective, communication serves as a social tool to

facilitate cooperation between agents, treated the same as actions in cooperation (Tomasello, 2010; Vygotsky, 1978). The context of the communication is the cooperative task itself. This perspective also adopts the Gricean cooperative communication axioms with the support from abundant evidence from comparative and developmental psychology.

Human cooperation has its unique nature. Humans show a motivation to help others from a very young age, while other animals do not understand this cooperative motivation (Fehr & Fischbacher, 2003; Tomasello, 2020). Take our close relative in evolution, chimpanzees as an example, they demonstrate sophisticated physical cognitive abilities (Barth & Call, 2006; Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007) and social cognitive abilities: They can understand others' goals (Warneken, Hare, Melis, Hanus, & Tomasello, 2007; Warneken & Tomasello, 2006; Yamamoto, Humle, & Tanaka, 2012) and intentions (Buttelmann, Carpenter, Call, & Tomasello, 2007; Tomasello, Carpenter, & Hobson, 2005), simulate and manipulate others' perception and knowledge (Hare, Call, Agnetta, & Tomasello, 2000; Melis, Call, & Tomasello, 2006), and understand false beliefs, the ability to interpret actions based on others' beliefs even when they contradict with reality (Buttelmann, Buttelmann, Carpenter, Call, & Tomasello, 2017). The chimpanzees show these capacities that we used to think unique to human beings, yet however, they still cannot cooperate. In the collaboration between chimpanzees, like their group hunting, each chimpanzee uses the others as social tools to achieve a greater individual reward (Tomasello, 2019). This reflects in two aspects. First, their group hunting strategy is the same as their individual hunting strategy (Bullinger, Wyman, Melis, & Tomasello, 2011): they do not have commitment to the task on themselves (Greenberg, Hamann, Warneken, & Tomasello, 2010) or expect commitment from others (Warneken & Tomasello, 2006). Second, they also do not share the reward with other chimpanzees that take supportive roles in the hunting. In hunting, they attempt to choose the most rewarding role. The dominant individual simply takes all the spoil after a collaborative effort (Melis, Schneider, & Tomasello, 2011). They only give a small portion of the spoil to other chimpanzees if they otherwise risk losing all the rewards

to those begging for spoil (John, Duguid, Tomasello, & Melis, 2019).

On the contrary, human toddlers show commitment and fairness in cooperation from a very early age. In cooperation, they keep their own commitment by continuing putting efforts into a joint activity even after receiving their own share of rewards to help their partners gain rewards (Hamann, Warneken, & Tomasello, 2012). When they break their commitment, they would acknowledge it (Gräfenhain, Behne, Carpenter, & Tomasello, 2009) and express guilt (Vaish, Carpenter, & Tomasello, 2016). They also want their partners to keep the commitment not to leave the joint activity when interrupted (Warneken & Tomasello, 2006). Children also show evidence that they see their collaboration with joint planning in which the roles of themselves and their partners are interchangeable (Carpenter, Tomasello, & Striano, 2005; Fletcher, Warneken, & Tomasello, 2012). They also share the rewards equally between roles (Hamann, Warneken, Greenberg, & Tomasello, 2011). It appears as if that they have a bird-eye view of the task (Tomasello, 2010).

The stark contrast between cooperation in humans and chimpanzees extends to communication. Human-raised great apes can even produce pointing-like gestures to invite humans to cooperate with them in obtaining food (Leavens & Hopkins, 1998). However, they are most likely using humans as a social tool in this process. The communication in great apes lacks the cooperative properties of human communication. Chimpanzees only understand pointing gestures from human to show the location of the food in a competitive setting (Hare & Tomasello, 2004), but not understand pointing gestures that show them the location of the food in a cooperative setting because they fail to understand the underneath helping intention (Tomasello, Call, & Gluckman, 1997). Apes are also not bothered a distracted or non-responding partner in an activity (Van der Goot, Tomasello, & Liszkowski, 2014). On the contrary, human infants communicate with pointing to cooperatively share interest with and helpfully provide information for a communicative partner(Liszkowski, 2009). Toddlers adjust their communicative attempts according to listeners' levels of comprehension (Golinkoff, 1993), communicate about absent but mutually known entities by pointing

(Liszkowski, Schäfer, Carpenter, & Tomasello, 2009), and infer the partner's intention and communicate to help (Moll & Tomasello, 2006; Tomasello & Haberl, 2003).

By comparing chimpanzees and children, the key to human cooperation and communication is the established jointness by the cooperation framework. In this cognitive perspective, the context of communication is the joint cooperative task. The same signal can be flexibly interpreted by human based on different joint tasks. In Liebal, Behne, Carpenter, and Tomasello (2009), 18-month-old infants and an adult cleaned up together by picking up toys and putting them in a basket. The adult stops and points to a toy. The infants can understand the context and pick it up and put it in the basket. However, when a second adult who has not shared this context enters the room and points to the same toy in the same way, the infants do not put the toy away into the basket because they did not share the same context. In another context where the infant and adult play stacking rings on a post, when the adult points to this same toy in the same way, infant bring the target toy back to the post. The jointness raised from different joint attentional frames provide different context in interpreting human communication signals.

1.3 Outlines of Successive Chapters

This dissertation aims to build a cognitive model framework for human communication, capturing its sparse, overloaded, and indirect nature. I take the Cooperative motivation perspective of communication and use joint tasks as contexts to define a mapping from (signal, context) to meaning called relevance. In Chapter 2, I first review in detail the fundamental building blocks for modeling human communication: mind, (joint) planning and inference. Human communication is modeled as actions over time that changes the receiver's mind. I review the computational progress in modeling the components of agent's mind: belief, desire, and intention. I also introduce Markov decision processes and partially observable Markov decision processes as the sequential decision-making frameworks the joint planning

process. Also, to the receiver, the communication can be seen as a piece of data or observation generated by the sender. Therefore, an inference of other agents' mind using the Bayes' rule is introduced. We will also introduce the formulation of the rational speech act model and the imagined we model. In Chapter 3, we show the advantage of human communication utilizing a joint task as context by using pointing as an example. I compare the pragmatic inference with joint perception and single agent perception of the same observation from the environment as a simulation of human communication and chimpanzee communication. I design a game called the guided Wumpus hunt, where a helper helping a player navigate in a sequential decision-making game to collect reward. The player navigates the world with partially observable Markov decision process as the decision-making model and Bayesian inference as the communicative model. Simulations show that agents communicate by asking why the signal is used receive higher rewards than agents interpreting the signal as its literal meaning. Details are summarized in Chapter 3 and published work (Jiang et al., 2021).

Respectively in Chapters 4, 5, and 6, I studied the three characteristics of human communication: overloaded, indirect and sparse nature. Chapter 4 focuses on the overloaded human communication. I derive the definition of relevance to summarizes the jointness-based context on top of how much the signal can help its receiver. In this chapter, I simulate extremely overloaded signals like the shout *ahh* in the guided Wumpus hunt, showing that a relevance-based communication model can flexibly and accurately interpret overloaded signals. Details are summarized in Chapter 3 and published work (Jiang et al., 2022).

Chapter 5 details a human behavioral study on the indirectness of human communication. I show that humans in fact use relevance when choosing between a set of utterances and prefer to cooperate with AIs that adopt relevance in communication. The results are also compared with GPT-4, the state-of-the-art large language model. Details are summarized in Chapter 3 and a manuscript prepared together with Boxuan Jiang, Anahita Sadaghdar, Rebekah Limb, and Tao Gao.

In Chapter 6, I study the problem of best timing for communication, reflecting the spar-

sity of human communication. A model based on Markov decision processes and partially observable Markov decision processes is derived, accompanied with a simulation study showing that relevance-based communication model works better than heuristics. In Chapter 7, we summarize the preceding chapters, discussed the theoretical and practical remarks on the relevance-based model, and proposed potential future research directions and applications.

CHAPTER 2

Modeling Mind, Decision-Making, and Inference

To our pets, to potential extraterrestrial creatures, writing to pen pals, prompting large language models, we, as humans, communicate every day. In 1974, the Arecibo message was sent to M13 globular cluster. From October 2020 to September 2021, nearly 71 trillion emails and 60 trillion spam messages were exchanged (Clissa, Lassnig, & Rinaldi, 2023). The sending and receiving process of a message is defined as communication in Shannon's communication model (Shannon, 1948). However, it is important to distinguish all information exchanges and communicative messages. When we use a calculator, we send information to the processor while the processor sends back the result of the calculation. Messages are exchanged between human and calculator, but it remains questionable whether this counts as an example of communication. What do we recognize in a communication message? Take a calculator and an AI assistant like Siri or Alexa as two examples, we are more likely to consider talking to an AI assistant as communication than pressing a calculator. The AI assistants are designed to process natural language and output natural language to user, which is more similar to everyday communication; they are regarded as more intelligent because it can flexibly handle a larger variety of tasks and can even learn from their interactions with the user. In a word, AI assistants seemingly own a mind. In this Chapter, we introduce how mind is modeled, how it influences our decision-making in communication, and how to infer a mind based on perceived actions and messages.

2.1 Decision-making with Mind

As human beings, we all have a model for how a mind works. This capacity to infer others' mental states and predict their future actions relies on Theory of Mind (ToM) abilities. ToM allows agents to infer others' beliefs as the informative state, desires as the motivational state, and intentions as the deliberative states of the mind (Bratman, 1987). Let's take the library example in Chapter 1 and analyze components of ToM. When Cindy retrieves the book from the shelves, her belief is that someone is waiting for the book. She may not know who it is, but she could have a guess. She also has multiple possible desires: she may want to deliver the book to the borrower; she may want to have lunch; she may want to go home and check on her daughter. Reconciling these diverging desires, she has the intention of delivering the book. Work in developmental psychology has shown that ToM is a social commonsense that develops in early infancy (Gergely & Csibra, 2003; Woodward, 1998; Wellman, 2018).

A famous example of ToM is the false belief task (Wimmer & Perner, 1983). In the false belief task, a child watches a display of a room accommodating two people, Sally and Anne, and two containers, a box and a basket. In the display, Sally puts a ball in the basket then leaves the room. Anne picks the ball up out of the basket and puts it into the box. When Sally returns to the room, where will she look for her ball? In this scenario, the child observer and Anne knows that the ball is in the box, while Sally believes that the ball is still in the basket. Children with ToM ability can distinguish Sally's mind from their own mind and predict Sally's action to be going to the basket.

ToM is also a growingly important topic in the field of artificial intelligence. If we review what we regard as artificial intelligence, we may recognize two categories of artificial intelligence. One is a language-based question answering machine that origins from Turing test. These machines take the language input from the user and provides an answer to the question in language. Calculators, as a prototype for computers, are an example of question answering machine. Recently, the development in computer vision has extended the input

and output in question answering machines beyond pure language. Search engines and some large language models can take in images or videos as input and output. The other is an action-based machine. These machines act based on a set of actions. For example, robots that can navigate through a map or pour a cup of coffee for you, or alpha Go, whose actions are which spot to position the next piece.

These two categories of artificial intelligence have some common features. First, these machines have flexible problem-solving abilities as if they had something that can be called a mind. We tend to think that these machines have their own representation of the world, have their desire, and their ability to logically reason about the mechanisms of the world. For example, large language models like GPT-4 are trained on social ToM data sets, and thus show a slight degree of ToM understanding (Sap, LeBras, Fried, & Choi, 2022). Second, they can be built based on probability theory. Language-based machines benefits from computational linguistics in predicting the next word with a probabilistic model; Action-based machines maintains a representation of states and rewards in the world and decide their actions by estimating the expectation of rewards potentially obtained from taking these actions. As an example, by adopting probabilistic theory of mind, computer vision-based AI agent can handle more uncertainty brought by human in the perceived environment (Zhao, Holtzen, Gao, & Zhu, 2015).

We can model ToM in an agent-based probabilistic framework. An agent's mind be modeled by their belief b and desire d. We simplify the model by leaving out diverging desires and thus intention. With ToM, an agent can predict other agents' actions a if it knows what they believe and what they want, which is their belief b and desire d. We represent this predicted action as a probability distribution $P(a|b, d)$. In utility theory, desire can be modeled as a utility function (Sutton & Barto, 2018). Here we use the belief-action value function $Q_d(b, a)$ to represent an agent's evaluation of the desirability (utility) of an action a based on its belief b. Humans are shown to adopt the principle of rational actions (Gergely, Nádasdy, Csibra, & Bíró, 1995): all else being equal, agents are expected to choose

actions that most efficiently achieve their desires which is to maximize their expected utility. We can formulate this decision-making as,

$$a^* = argmax_a Q_d(b, a). \tag{2.1}$$

To capture a certain degree of stochasticity and irrationality in human nature, a softmax function can be used to calculate the probability,

$$P(a|b, d) \propto \exp\{\alpha Q_d(b, a)\}, \tag{2.2}$$

where α is a parameter of the agent's rationality. If α is high, the agent is more likely to choose actions with higher utility, which we consider more rational. If α is as low as 0, then the agent does not consider utility when choosing actions, which is less rational.

Communication is also an important part of human physical and social interaction. We can also regard the communication as an action driven by the agent's belief and desire to communicate, to reflect the causal relationship from mind to communication, from meaning to message. When a person wants to communicate, they must have some information by sharing which they can benefit. ToM plays an important role in human communication from a very early age. Communication is used as an effort for agents to increase their mutual benefits through synchronizing their minds. Toddlers aged 18-36 months consider others' perspectives, exhibiting greater amounts of pointing when their partner's view is obstructed (Franco & Gagliano, 2001). When toddlers play with caregivers, they are aware of whether a toy is blocked from the caregiver's sight and use different pointing strategies accordingly (Franco, 2005). More than that, when children observe the actions or gestures from an animated agent, the children can infer the goal of the agent, a component in the agent's mind (Woodward, 1998). In the following section, I will review a model of inferring the mind from actions with the Bayes' rule.

2.2 Infer the Mind: Bayesian Theory of Mind

We see how humans act as they make their decisions from what they believe and what they desire. We can also reverse this planning process to infer people's mind from their observed actions. A ToM agent can make sense of other agents' actions a from their belief b and desire d. For example, by observing a food-seeking person pass by the Mexican food truck to approach the Korean food truck, we are more likely to infer that the person prefers the Korean food truck to the Mexican food truck (C. Baker, Saxe, & Tenenbaum, 2011). If we have a set of potential beliefs B and desires D of an agent and observe its action as data, we can infer its belief b and desire d from its action a by reversing the planning process using the Bayes' rule,

$$P(b, d|a) \propto P(a|b, d)P(b, d). \tag{2.3}$$

Humans are Bayesian observers of other people's minds. With beliefs and desires as components of the mind, agents are capable of decision-making, perspective-taking (Barnes-Holmes, McHugh, & Barnes-Holmes, 2004), inferring other agents' minds based on their actions, changing beliefs according to their sensory input, and revising their plans. Bayesian ToM has also been successfully formulated as inference in inverse planning. Bayesian ToM successfully explains both infant (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Liu, Ullman, Tenenbaum, & Spelke, 2017) and adult (C. Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017) cognition.

Bayesian ToM is successful in modeling physical and social goal inference with a sequential decision-making model (C. Baker et al., 2011; Ullman et al., 2009). In C. Baker et al. (2011), human participants are a shown an agent navigating in a grid world w with several possible goals G. Each goal g in the world is modeled as a reward structure of the agent. The goal can be seen as a desire, as it represents the agent's reward structure, or can be seen as a belief, because it represents the agent's knowledge of where the goal is. The agent owns a goal $g^* \in G$, takes a sequence of actions $a_{1:T}$ over time following the principle of rational actions,

resulting in a trajectory $s_{1:T}$. In these models, the participants observe the environment, trajectories or the social interactions from the outside as an observer. They do not engage as a part in the social interaction or provoke communication. Assuming that the agent's goal is each possible goal g, the participants simulate the agent's actions and trajectories,

$$P(s_{1:T}|g, w) = \sum_{a_{1:T}} P(s_{1:T}|a_{1:T}, w)P(a_{1:T}|g, w). \tag{2.4}$$

Then the participants use the observed trajectory infer the actual goal of the agent with Eq. (2.3),

$$P(g|s_{1:T}, w) \propto P(s_{1:T}|g, w)P(g|w). \tag{2.5}$$

2.3 Sequential Decision-making Models

2.3.1 Markov Decision Process

Markov decision processes (MDPs) provide a framework on how an agent interacts with the environment. In this model, the agent and environment interact at each of a sequence of discrete timesteps, $t = 0, 1, 2, 3, \cdots$. The environment is represented as in different states, that could possibly transition to each other. The agent has a collection of actions. In each timestep, the agent receives a representation of an environment state and chooses an action based on the state; then the environment responds to the action by presenting to the agent a new representation of state and a reward which the agent chooses actions to maximize over time (Sutton & Barto, 2018).

As its name suggests, an MDP has the Markov property: the state transition at time t only depend on the state and action at time t-1. In other words, "the state must include information about all aspects of the past agent–environment interaction that make a difference for the future" (Sutton & Barto, 2018). An MDP can be described as a tuple (S, A, T, R). In the tuple, S is the state space, a set of states of the world. A is the action space, a set of actions. $T : S \times A \to \Delta_S$ is the transition function measuring how likely the agent is

transitioned to the next state s' by starting from the current state s and taking action a,

$$T(s, a, s') = P(S_t = s' | S_{t-1} = s, A_{t-1} = a). \tag{2.6}$$

Here Δ_S represents the set of probabilistic distributions over S. $R : S \times A \times S \to \mathbf{R}$ is the reward function, which provides the immediate reward gained by the agent if it starts from state s, takes the action a, and ends in state s'.

$$R(s, a, s') = E(R_t | s_{t-1} = s, A_{t-1} = a, S_t = s'). \tag{2.7}$$

In some literature, the initial state s_0 is included in the tuple (Kearns, Mansour, & Ng, 1999).

In MDPs, the agent chooses actions to maximize the expected long-run reward. A policy is a rule used by the agent to choose actions from state information. In an MDP, depending on the maximum time steps of the agent-environment interaction called horizons, a policy has respective representations in two frameworks. One framework is finite-horizon optimality, where the agent maximizes the reward to be received in the next finite T steps, which can be represented as $E(\sum_{t=0}^{T-1} r_t)$. In a finite-horizon optimality framework, an agent can choose actions based on a non-stationary policy, which is a sequence of state-action mapping indexed by time, or a stationary policy, which is a state-action mapping, which does not change by time. For example, if you are in a last-minute shopping for Valentine's Day flowers or if you still have two days to order flowers, your policies are very likely to be completely different. An example of non-stationary policy is a policy tree, which we will introduce in Chapter 2.3.4.

The other framework is the infinite-horizon optimality where the agent maximizes the discounted reward to be received in the whole future, which can be represented as $E(\sum_{t=0}^{\infty} \gamma^t r_t)$. γ is called the discount factor. In an infinite-horizon optimality framework, an agent chooses actions based on a stationary policy $\pi(a|s)$, a probabilistic distribution over actions given a current state s.

In this dissertation, our formulation of policy is written as stationary policies $\pi(a|s)$. However, the formulation of policy prediction (Chapter 3), policy evaluation (Chapter 3), and relevance (Chapter 4) can also apply to non-stationary policies.

To solve an MDP is to find the optimal policy, which maximizes the agent's expected long-run reward. Dynamic programming is a widely used solution to MDP. In a finite-horizon optimality framework, the optimal policy can be provided by dynamic programming the value of the initial state $V(s_0)$, which is the maximum expected long-run reward an agent can get by following any policy starting from the state. For example, in a k-horizon task, the value of a state s can be calculated by the Bellman update,

$$V_{k+1}(s) = max_a \sum_{s'} T(s, a, s')(R(s, a, s') + V_k(s')), \tag{2.8}$$

where $V_k(s')$ is the value of state s' in the $(k-1)$-horizon task. The optimal non-stationary policy is the action $a^* = argmax_a \sum_{s'} T(s, a, s')(R(s, a, s') + V_k(s'))$ followed by the optimal policy of the $(k-1)$-horizon task. The optimal policy of the 1-horizon task is $\pi_1^*(s) = argmax_a \sum T(s, a, s')R(s, a, s')$.

In infinite-horizon optimality framework, value iteration is widely used to provide the optimal policy. With the infinite-horizon version of Bellman update,

$$V_{k+1}(s) = max_a \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V_k(s')), \tag{2.9}$$

the sequence of value estimates for state s, $(V_0(s), V_1(s), V_2(s), \cdots)$, converges to $V(s)$, which is the value of the state.

We can use the state-action value function $Q(s, a) = \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V(s'))$ to represent an agent's expected future reward by taking an action a from a state s. Adopting the principle of rational actions, the policy of the agent is $\pi(a|s) \propto \exp\{\alpha Q(s, a)\}$.

In this dissertation, we adopt value iteration to solve MDPs.

2.3.2 Partially observable Markov decision process

In MDPs, the agent knows the state of the world s. A partially observable Markov decision process (POMDP) is an MDP in which the agent cannot observe the current state directly. Instead, the agent interacts with a partially observable environment, and makes observations to help estimate the current state (Kaelbling, Littman, & Cassandra, 1998).

A POMDP can be described as a tuple (S, A, Ω, T, R, O). Same as in an MDP, S is the state space, A is the action space, T is the transition function, and R is the reward function. As an addition, Ω is the observation space, a set of possible observations that is collected through actions from states; $O : S \times A \to \Delta_\Omega$ is the observation function which describes the agent's sensory process.

$$O(s', a, o) = P(O_t = o | S_t = s', A_{t-1} = a), o \in \Omega. \tag{2.10}$$

It shows the probability that an agent gets the observation o if it ends in the state s' by taking the action a. For example, if a person does not know which hemisphere he is in, he can take the action of looking up at night. If he is in the southern hemisphere, he is more likely to see the observation of the Southern Cross; If he is in the northern hemisphere, he is more likely to see the Polaris.

2.3.2.1 Belief state

In a POMDP, the agent cannot observe the current state directly, but estimates the current state. This estimate b called a belief state, is a probability distribution over all states. $b(s) = P(S = s), s \in S$ shows the probability of the current state being state s.

The belief state can be updated by the agent's observations. In one timestep of POMDP, the agent receives an observation o by taking an action a. The agent can update its belief state b with the new information using the Bayes' rule:

$$b'(s') = P(s'|b, a, o) = \frac{P(o, s'|a, b)}{P(o|a, b)} = \frac{O(s', a, o)T(s, a, s')b(s)}{\sum_{s'} O(s', a, o)T(s, a, s')b(s)}. \tag{2.11}$$

We can also write down this belief transition process as

$$b' = SE(b, a, o) \text{ if } b'(s') = P(s'|b, a, o), \forall s' \in S, \tag{2.12}$$

where SE is short for state estimator, representing the belief after the agent takes action a and obtains observation o. As shown in the belief state update, the belief state in step $t+1$ only depend on the belief state in step t, the action in step t, and the observation in step $t+1$, which means the belief state framework has Markov property: A POMDP is an MDP whose state space is the belief space of the POMDP.

The belief space is the set of all possible belief states, which we can represent as ΔS. A belief is a point in the belief space of a POMDP, also called a belief state. The belief-space MDP can be represented as a tuple $(\Delta S, A, \tau, \rho)$. τ represents the transition function between beliefs:

$$\tau(b, a, b') = \sum_o I(b' = SE(b, a, o) \sum_{s'} O(s', a, o) \sum_s T(s, a, s') b(s), \tag{2.13}$$

where I is the indicator function, being 1 if $b' = SE(b, a, o)$ and 0 otherwise. ρ represents the reward function between beliefs:

$$\rho(b, a, b') = \sum_o I(b' = SE(b, a, o)) \sum_s b(s) \frac{\sum_{s'} T(s, a, s') O(s', a, o) R(s, a, s')}{\sum_{s'} T(s, a, s') O(s', a, o)}. \tag{2.14}$$

2.3.2.2 POMDP agent policy

Same as MDPs, POMDPs can be categorized as finite-horizon and infinite-horizon POMDPs. In a finite horizon POMDP problem, the agent selects policy to maximizes the expected long-run reward in the next T steps $E(\sum_{t=0}^{T-1} r_t)$, and the non-stationary policies are a policy trees. A 1-step policy tree is an action. A t-step policy tree is a policy tree that starts from an action a as the root, spreading a branch with each possible observation on which a $(t-1)$-step policy tree is attached. When an agent follows a t-step policy tree, it performs the action in the root, receives an observation, and follows the $(t-1)$-step policy tree on the branch of the received observation. In an infinite horizon POMDP problem, the agent

selects policy to maximize the expected long-run reward $E(\sum_{t=0}^{\infty} \gamma^t r_t)$, and the policies are stationary policies, a mapping from a belief to a probability distribution of actions $\pi(a|b)$.

Starting from a belief state b, we can calculate the expected reward $Q(b, p)$ that the agent can receive from following a policy p. This is achieved by calculating the expectation of expected reward obtained from executing the policy on each state $Q(s, p)$,

$$Q(b, p) = \sum_{s} b(s)Q(s, p). \tag{2.15}$$

Therefore, on belief state space, the value function of a certain policy $Q(b, p)$ is linear. It is the inner product of the belief state b and a vector $\alpha_p = (Q(p, s_1), Q(p, s_2), \cdots)^T$. Therefore, each policy can be represented by its state-value vector α_p.

The maximum expected reward $V(b)$ the agent can receive starting from belief state b is the maximum value of all possible policies at belief state b. $V(b) = max_{p \in P} Q(b, p)$. If the agent has t steps left, the policy set P contains all the t-step policy trees, which are finite in number. The value function can be written as $V_t(b)$. Since the t-step value function is the maximum of a finite number of linear functions $V_p(b)$, the value function for a t-horizon POMDP is piecewise linear and convex. Each linear piece of t-step of the value function is a policy that is optimal for some belief states. For infinite-horizon POMDP problems, the policies are not policy trees but mappings from a belief to a distribution of actions. In this scenario, the value function $V(b)$ is still convex but not piecewise linear because it is the maximum value of infinite many policy trees.

2.3.3 POMDP Solutions

A POMDP is far from easy to solve. For a t-horizon POMDP problem, a direct solution is to enumerate all t-step policy trees and compare their policy value function. At each belief state b, the t-step policy tree that has the maximum policy value is the optimal t-step policy. $p^* = argmax_{p \in P_t} Q(b, p)$. For infinite-horizon POMDPs, a common practice is to solve with value function $V(b)$. However, value function is difficult to calculate. Many approximation

methods of solving POMDP attempt to approximate the infinite horizon value function. Another way to solve infinite-horizon POMDP problems is to approximate with the solution of high horizon POMDP problems. In high horizon, the value function and the root of the policy tree barely change. In this section, we introduce a few POMDP solvers that we will utilize in the following chapters.

2.3.3.1 V_{MDP} and Q_{MDP}

A naïve way to solve POMDP is to solve the underlying MDP in the POMDP framework assuming full observability. The V_MDP method solves the MDP and get the value of each state $V_{MDP}(s)$, then approximate the value function as $V(b) = \sum_s b(s)V_{MDP}(s)$ (Cassandra, Kaelbling, & Kurien, 1996).

Q_{MDP} is a slightly more sophisticated method than V_{MDP}. It calculates the Q value $Q_{MDP}(s, p)$ for each state-action pair in the underlying MDP. The Q value $Q(b, p)$ for each belief-action pair is calculated as $Q(b, p) = \sum_s b(s)Q_{MDP}(s, p)$. The stationary policy is $p^* = argmax_a Q(b, a)$ (Littman, Cassandra, & Kaelbling, 1995).

V_{MDP} and Q_{MDP} are very efficient but not accurate approximation methods.Agents following them do not take actions that only gather information. For POMDPs whose believes do not change based on observations, these methods can give accurate estimations to the value functions. We will use these methods in Chapter 6.

2.3.3.2 Grid-based methods

Grid-based solution to POMDP uses a grid of belief states as a kernel to estimate the value function. A grid-based method requires initializing a grid, grid point evaluation, and interpolation.

A grid can be defined as $G = \{b = \frac{m}{M} | m \in I_+^{|S|}, \sum_i m(i) = M\}$, which is called a fixed-resolution regular grid. M is a positive integer called the resolution of the grid and $I_+^|S|$ is

a vector of positive integers whose dimension is the cardinality of the state space (Lovejoy, 1991). For example, a 3 dimensional grid with a resolution of 2 is $\{(1,0,0), \left(\frac{1}{2},\frac{1}{2},0\right), \left(\frac{1}{2},0,\frac{1}{2}\right), (0,1,0), \left(0,\frac{1}{2},\frac{1}{2}\right), (0,0,1)\}$.

With the grid defined and an arbitrary value function initialized, we can start the value iteration with Bellman update for each $b \in G$:

$$V(b) = max_a \sum_o \tau(b,a,SE(b,a,o)) \left[\rho(b,a,SE(b,a,o)) + \gamma V(SE(b,a,o))\right] \tag{2.16}$$

Then the belief states off the grid are evaluated by a weighted sum of points in the grid in the interpolation step.

2.3.3.3 Point-based value iteration

Like grid-based method, point-based value iteration (PBVI) solution to POMDP uses a set of belief states as a kernel to estimate the value function. A PBVI consists of two parts: value iteration and belief expansion (Pineau, Gordon, Thrun, et al., 2003; Ross, Pineau, Paquet, & Chaib-Draa, 2008; Shani, Pineau, & Kaplow, 2013).

For each belief in the kernel $b \in B$, the value is represented by its optimal policy $\alpha^*(b)$. In each step of the value iteration, the α vectors are updated with the backup operator. The new policy should start from an action $a \in A$, followed by an optimal policy α for each $SE(b,a,o)$. For each action a and observation o, the optimal policy is

$$\beta_{(a,o)} = argmax_\alpha(\alpha^T SE(b,a,o)). \tag{2.17}$$

Then if we start from a state s, by taking the action a, we are expected to obtain the reward

$$\beta_a(s) = \sum_{s',o} \frac{O(s',a,o)T(s,a,s')}{\sum_{s'} O(s',a,o)T(s,a,s')} \left[R(s,a,s') + \gamma\beta_{(a,o)}(s')\right] \tag{2.18}$$

Therefore, the action a with the largest is selected for the root of the new optimal policy:

$$\alpha = argmax_{\beta_a} \beta_a^T b \tag{2.19}$$

With the updated policies, the values of the beliefs off the grid are interpolated by $V(b) = \max_{\alpha} b^T \alpha(b'), b' \in B$.

Since each belief state in the belief set B provides an α-vector for approximation to the value function, the larger B is in size, the more accurate the approximation is. After each value iteration converges, we can expand the belief set. In PBVI, the value iteration and the belief expansion are performed alternately and can be stopped at any time. The PBVI algorithm is an anytime algorithm. If stopped early, fewer belief points are in the belief set, thus giving a less accurate approximation but saves computational time and resources; if stopped late, more belief points are in the belief set, thus giving a better approximation but cost more time and resources (Shani et al., 2013).

2.3.3.4 PERSEUS

PERSEUS is an approximate point-based value iteration algorithm for POMDPs (Spaan et al., 2006). It keeps the initiation and expansion components of PBVI but proposed a more efficient value function update. In one iteration in the improve step of PBVI, the value of each belief in the belief set B is updated through the backup procedure. In one iteration in the improve step in PERSEUS, one belief $b_0 \in B$ is selected. Its value $V(b_0)$ and policy vector $\alpha(b_0)$ are updated through the backup procedure. Then the policy vector is evaluated with each $b \in B$ as $Q(b, \alpha(b_0))$. If the new evaluation $Q(b, \alpha(b_0))$ is higher than the not-yet updated value $V(b)$, then the belief b is removed from B in this iteration because a better policy has been found for b. The iteration of improvement ends when each $b \in B$ is updated or removed. Because much fewer beliefs go through backup, the improvement procedure in PERSEUS is very efficient. I will use PERSEUS to solve POMDPs in Chapters 5 and 6.

2.3.4 Multi-agent POMDPs

In a multi-agent setting, the state of the world is less observable and predictable. Not only the physical world may be partially observable to agents, but the agents also cannot access the other agents' minds. Several POMDP-based models have been proposed to model the multi-agent decision-making process.

2.3.4.1 Decentralized POMDPs

The decentralized POMDP (Dec-POMDP) framework models cooperative interactions of humans or artificial intelligent agents in a partially observable environment. It is a model of multiple agents making decisions in uncertainty without a central controller making decisions on-the-fly (Oliehoek, Amato, et al., 2016). It can be defined as a tuple $(D, S, A, \Omega, T, R, O, h, b_0)$. $D = \{1, 2, 3, \cdots, n\}$ is the set of id of agents; A is the joint action space for agents; Ω is the joint observation space for agents; h is the horizon of the problem; the other components have the same definition as in the POMDP framework.

Here we take a two-agent example to illustrate the dynamics of Dec-POMDP. Two agents take actions based on their individual belief. The agents do not know other agents' actions. The environment states transition based on their joint action, and provide an observation of which each agent can only see a part. The reward is non-observable by either of the agents. The agents use their individual action and observation to update the belief and takes another action. That is to say, the policy of each agent is a function of its own action-observation sequence $(a_1, o_1, a_2, o_2, \cdots, a_{t-1}, o_{t-1})$. For example, two lunar roving vehicles are deployed on the moon. Their policies are like: if you see a rock, the go past it; if you see the station, then go get supplies and wait until you see the other vehicle; if you see the other vehicle, start retrieving dirt samples.

Even though Dec-POMDP agents make decisions without a central controller, their policies are solved in a centralized way. Before timestep 0, a central controller plans for each

agent. For the lunar rover example, the central controller considers all the action-observation sequences of the two agents even before they were deployed to the moon: If Vehicle 1 moves west and observes observation *station* while Vehicle 2 moves south a' and observes observation *station*, then both of the vehicles should move to the station.

However, since the agents do not share observations when executing the policy, the policy needs to be executed in a decentralized manner. If an agent sees the action and observation sequence $(a_1, o_1, a_2, o_2, \cdots, a_{t-1}, o_{t-1})$, what should the agent do in timestep t is predetermined by the central planner. After the policy for each individual agent is fixed, the agents start the navigation and act according to its received sequence of action-observations. In the lunar rover example, the two vehicles execute their own policy individually without considering the other vehicle: I am going to the station no matter what the other vehicle does. This is an example of two agents both do something parallely, but not jointly do something together.

2.3.4.2 Interactive POMDPs

Interactive POMDP (i-POMDP) is also a model of multi-agent making decisions under uncertainty (P. J. Gmytrasiewicz & Doshi, 2005). However, different from Dec-POMDP, it models both planning and execution in a decentralized manner. The i-POMDP takes the perspective of one agent and treat other agents as a part of uncertainty.

An interactive POMDP can be defined as a tuple (IS, A, Ω, T, R, O). $IS = S \times M_j$ is the interactive state space, which is a set of physical states and the models of other agents. The models of other agents include its history of observations, policy and observation function, and may have two variants: sub-intentional model and intention model. An agent with sub-intentional model treats other agent as a noise in the POMDP environment, while an agent with intentional model infers other agents' models. When there is no evidence showing that an agent is using the intentional model, the sub-intentional model is applied. For example, in a rock-paper-scissors game, if you do not see any evidence that the opponent studied your

pattern of play, you can assume that they do not have any strategies planned against you.

An interactive POMDP is a POMDP on the interactive state space. The agent that follows a interactive POMDP takes an action based on its belief on both the physical state and the other agents' models. Getting an observation from the environment, the agent updates its beliefs on both the physical state and the other agents' models based on the observation and takes an action that maximized the expected long-run reward.

The solution to interactive POMDPs is to maintain explicit models of the other agents. By solving other agents' possible models, the agent can better predict their actions. Once an agent has a good prediction of the other agents, it can plan its action accordingly to maximize the expected long-run reward. For example, in a rock-paper-scissors game, if you have a model of the other agent to be more likely to play the paper, then you should play the scissors to gain a higher expected reward.

This planning process involves a recursive nature (Doshi, Gmytrasiewicz, & Durfee, 2020). A sub-intentional model is the entry level of the recursion. For example, if you do not see any evidence that the opponent studied your pattern of play, you can assume that they do not have any strategies planned against you, then you can use your knowledge of their playing style to build a strategy to win over your opponent. Interactive POMDPs are very difficult to solve, with a high-dimensional recursive interactive state space.

Like decentralized POMDPs, interactive POMDPs also provide a framework which may be helpful with studying the interactions between agents in partially observable environments. As we have already seen from the examples, interactive POMDPs can model multi-agent interactions including both cooperation and competition.

2.4 Models of Communication as Decision-making

After reviewing how mind, decision-making, and inference are modeled, I start to review how these components scaffold the models of communication.

2.4.1 Rational Speech Act

One ToM-based communication model is the rational speech act (RSA) model (Frank & Goodman, 2012; Goodman & Frank, 2016). RSA was developed to solve reference games. In a reference game, a target exists among a set of possible targets, for example, the blue square among W=green square, green circle, blue circle. A speaker selects one signal from a set of linguistic utterances, for example, C=*green*, *blue*, *square*, *circle*, and sends the signal to a listener trying to help the listener identify the target. In RSA, each utterance is treated as an action with a utility function. The utility of an utterance is defined by how the utterance is expected to change the listener's belief to reflect the target. The generation and interpretation of an utterance can be modeled with the principle of rational actions: maximizing the expected utility.

In RSA, the listener's belief is modeled as a probability distribution over all possible targets $P(w)$. The belief of the listener L changes upon receiving each utterance u can be represented as $P_L(w|u)$. The speaker S estimates the utility of each utterance by how much the listener's belief is consistent with the target with a utility function $U(u;w) = \log P_L(w|u)$. Then assuming that the speaker chooses utterances rationally, if the target is w, the probability that she chooses the utterance u is:

$$P_S(u|w) \propto \exp\{\alpha U(u;w)\}. \tag{2.20}$$

Meanwhile the listener guesses the target with a Bayesian inference, also by estimating the utility of the signal,

$$P_L(w|u) \propto P(w)P_S(u|w). \tag{2.21}$$

The recursion starts from a model of a literal listener who interprets the signal for its literal meaning. $P_{Lit}(w|u) \propto \delta_{[u](w)}P(w)$, where $\delta_{[u](w)} = 1$ if w is consistent with the literal meaning of u, and 0 otherwise. For example, the target is the green circle. Before any utterance is sent, the listener's belief is $\frac{1}{3}$ for each possible target. Upon receiving an utterance *green*, a literal listener will assign 1 to the green square and the green circle

and thus for these two w, $P_{Lit}(w|u) = \frac{1}{2}$. He will also assign 0 to the blue circle and its $P_{Lit}(w|u) = 0$.

Since the target is the green circle, the literal speaker, accounting for the belief change of the literal listener, evaluates the utterance *green* to have a utility of $\log \frac{1}{2}$. The utterance *circle* has a utility of $\log \frac{1}{2}$, *blue* has a utility of $-\infty$, and *square* has a utility of $-\infty$.

The literal speaker has a $\frac{1}{2}$ probability to send out *green*, and $\frac{1}{2}$ probability to send out *circle*. Then upon receiving the utterance *green*, the pragmatic listener who considers how the literal speaker evaluates the utility of each utterance, and the probability of sending them: If the target is the green square, then the utility of *green* is $\log \frac{1}{2}$, while the utility of square is $\log 1$ and the utility of blue or circle is $-\infty$. Assuming that $\alpha = 1$ in equation, the probability of sending out green is $\frac{1}{3}$. If the target is the green circle, then the utility of green is $\log \frac{1}{2}$, while the utility of circle is $\log \frac{1}{2}$ and the utility of blue or square is $-\infty$. The probability of sending out green is $\frac{1}{2}$. If the target is the blue circle, the probability of sending out green is 0. Weighing these probabilities, the pragmatic listener will have $P_L(\textit{green circle}|\textit{green}) = \frac{3}{5}$, and $P_L(\textit{green square}|\textit{green}) = \frac{2}{5}$. Through this recursion, the listener prefers the green circle over the other two possible targets. Intuitively, this results from comparing the utterance with the context: what has been said and what could have been said.

RSA has been shown successful in modeling a variety of phenomena in linguistics, including metaphor (Kao, Bergen, & Goodman, 2014), redundancy (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020), and convention formation (Hawkins, Frank, & Goodman, 2017). However, as we discussed in Chapter 1, RSA is highly dependent on the literal meaning of the utterances. For example, in a reference game, when I say *dance*, the target is more likely to be a ballerina than a policeman as suggested by the utility function. However, if we expand the task beyond reference game or restrict the utterance space, the literal meanings of the utterances may not be consistent with the targets and thus the strategy may change. There are also some communicative signals that do not have their literal meaning. The mechanism

of deploying and interpreting these signals may exceed the scope of RSA.

Multiple works attempt to extend RSA to action-based agent interaction with physical environments. When people want to show their goal to an observer, they will take actions different from what they would do individually (Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016; Ho, Cushman, Littman, & Austerweil, 2021). Humans can also infer the communicative intention from actions that are inefficient in utility, especially repetitive actions (A. Royka, Chen, Aboody, Huanca, & Jara-Ettinger, 2022; A. L. Royka, Török, & Jara-Ettinger, 2023). I will also take the perspective that communication is an action, and we think the goal of the communicative action is to increase the joint utility. I will elaborate in Chapter 4.

2.4.2 Communicative Interactive POMDP

Communicative IPOMDPs (CI-POMDPs) (P. Gmytrasiewicz, 2020) build on interactive POMDPs but introduces a communication channel. A CIPOMDP for an agent i is defined as a tuple $(IS_i, A_i, M_i, \Omega_i, T_i, R_i, O_i)$, where M_i contains all the messages that can be sent by the agent. In each timestep, the agent can take two actions: one physical action $a_i \in A_i$ and one message $m_i \in M_i$. It also receives two observations: one is generated from the state and the other is the message generated from the other agent. The agent considers how the message sent and received changes the utility of the partner and the agent itself.

In CI-POMDP, how the messages changes the beliefs are not directly provided, which makes the model difficult to apply on actual communication utterances. Also, the communication in CI-POMDP lacks the transparency brought by the jointness in human communication. The defined interactive states include the whole model of the partner, making the complexity of the model drastically increase as agents recursively reason about each other.

In this dissertation, I aim to build communication models for overloaded, indirect, and sparse signals. Like in RSA and CI-POMDP, the communicative signals are defined as

rational actions by the sender, and observations by its receiver. I take advantage of the transparency of human communication to set constraints to simplify the model. I also propose how agents exchange and coordinate their minds within the joint framework of communication. In the next chapter, we start from an example of visual communication between pure communicative messages and pure actions showing communicative intentions, between having a single literal meaning and having no literal meanings, pointing, to illustrate our framework of human communication and mind coordination.

CHAPTER 3

Communicate via Paternalistic Jointness

3.1 Pointing as Sparse, Overloaded, and Indirect Visual Communication

Imagine two hunters hunting in a forest. The young hunter sees a broken branch on the ground. Assuming that the branch was broken by the wind, he gets ready to continue his search. At this moment, his partner, the experienced hunter points the broken branch to him. The young hunter suddenly realizes that the branch was broken by their prey. He holds his breath and prepares to hunt.

Human pointing is a representative example for overloaded, indirect and sparse human visual communication. This pointing gesture is sparse; the semantics of such a succinct act has complex meanings, much beyond the gesture itself: "Take a look at this broken branch caused by a prey." Such rich information is condensed spatially and temporally into one extension of the index finger, which lasts no more than a few seconds. Second, pointing is overloaded. Multiple features may coexist in the location where a pointing signal directs, and each of these features may be the referent of the pointing. Third, pointing is indirect. The meaning of the pointing can go far beyond the referent visual feature. In the hunting example, the experienced hunter simply points to the broken branch, but she does not mean the brokenness of the branch. Instead, she means they should get ready to hunt.

Fortunately, humans can accurately interpret pointing gestures (Kendon, 2004; Lascarides & Stone, 2009) despite their key properties of being sparse, overloaded, and indirect.

As a rich form of communication, pointing is effective in changing its receiver's mind and actions.

Until Wittgenstein called attention to its underlying complexity, pointing had generally been perceived as an intuitive, unremarkable communicative gesture. It is among the most conspicuous and common forms of human communication. Children point to help adults retrieve objects they were looking for, foraging partners point out the potential locations of food to help each other, and customers point to their empty glasses to request assistance from the server. Pointing is among the first communicative gestures human infants learn to use (Butterworth, Simion, et al., 2013): at as young as the age of one, infants use pointing to communicate information (Liszkowski, Carpenter, Striano, & Tomasello, 2006). Although pointing is pervasive in everyday human life, it is rarely observed in wild animals. Human-raised great apes can produce pointing-like gestures to invite humans to cooperate with them in obtaining food (Leavens & Hopkins, 1998), yet pointing in great apes lacks the cooperative properties of human pointing: apes are not bothered when the partner is distracted or non-responding (Van der Goot et al., 2014). Chimpanzees' failure to achieve a deep understanding of pointing suggests that human pointing may reveal surprising intricacy of human communication.

In the above pointing example, one's attention to an observation is not a spotlight to enhance individual sensory accuracy and more than an action label commonly adopted in the computer vision community. Instead, it involves rich cognitive inference and demonstrates properties of human unique communication. Instead, he should try to make sense of why the sender attracts his attention to the referent. When an agent attends to his surroundings *individually*, he owns his observations; it is his own job to evaluate the relevance of perceived objects and filter out irrelevant information (Wilson & Sperber, 2006). In contrast, when a signaler points an object to a receiver, she invites the receiver to become a "guest" to the observation. The pointing gesture creates a triangulation in joint attention between the sender of the pointing, the receiver of the pointing, and the referent (Brinck, 2004; Eilan,

2005). When interpreting the signal, the receiver does not think about why the referent attracts his attention, but why the sender invites him to pay attention to the referent. The receiver can safely assume that the information given by the "host" must be relevant to the shared task. In the hunting example, no matter how clearly the young hunter sees the broken branch individually, he is not likely to change his explanation of the shape of the branch. However, from a point, he can make much richer inferences and revise his plans more drastically. Both the signaler and the receiver of the pointing are aware of the effect of this joint attention, so they reserve it to convey relevant information.

A crucial function of pointing is helping. Crucially, it is a particular type of helping with two unique characteristics. First, the helper is in a position to help because her belief is closer to reality, not because she has any physical advantage. Therefore, pointing must involve diverging beliefs in which the helper knows better how to improve the partner's well-being. This type of helping is called paternalistic helping or Smithian helping in developmental psychology (Martin, Lin, & Olson, 2016). Second, unlike instrumental actions, pointing does not change the physical states at all. Its only function is to change the partner's mind. Therefore, models of instrumental helping would fail to apply (Ullman et al., 2009). Instead, we argue that it should be understood as an "utterance", a communicative action with a utility function in rational speech act (RSA) (Frank & Goodman, 2012). The generation and interpretation of an utterance can be modeled with the principle of maximizing expected utility from decision theory.

3.2 Paternalistic Helping

The concept of paternalistic helping is based on Adam Smith's discussion of empathy. During his discussion, he compared two types of empathy. First, he addressed Hume's definition, which proposes that empathy is a simple resonance of other's feelings. Hume's conventional definition of empathy proposes that empathy is the direct mirroring of another's mindset

(Hume, 1751/2018). In other words, the agent and subject involved in a particular act of empathy should share the same mindset. To better understand the two competing views addressed in Smith's argument, one can imagine a theoretical example involving you and your friend. Both of you are backstage preparing for your friend's performance in the school talent show. While your friend is excited to perform, you dread the performance because you know that he is objectively bad at singing. In this example, you and your friend each have a different perspective in regards to the same action of your friend performing. However, according to Hume's definition, to successfully empathize with them, you must abandon your personal opinion and take on your friend's perspective.

In contrast to this conventional definition, Smith proposed that true empathy involves the coordination of mindsets between the empathizer and the agent being empathized with, which is the paternalistic empathy. His own proposed definition of empathy is the act of investigating how an agent would feel if they, in their current state of consciousness, were placed into the target individual's situation (Smith, 1759/2010). In the talent show example, when executing the paternalistic empathy, you maintain your own mindset when evaluating the action of interest: If I were about to perform on the show and I were bad at singing, I would feel very bad. In this definition, empathizing involves applying your own mindset to evaluate the situation of your friend performing. Because you know your friend is bad at singing, your evaluation of how you would feel in their situation leads you to be worried that your friend may embarrass himself on stage. As illustrated in the example, Paternalistic empathy is beyond directly copy another person's mindset, but involves coordinating diverging mindsets of two agents. Your act of worrying is an attempt to coordinate your friend's belief of excitement about the upcoming performance with your opposing belief that their performance will have a bad outcome.

Paternalistic empathy has been explicitly extended to the well-studied phenomenon of human behavior known as paternalistic helping. Paternalistic helping involves a helper doing what she thinks to be good to the helpee, even if that is not what the helpee wants (Martin

et al., 2016). This act involves the coordination of two mindsets as the helper must balance the helpee's desire with what she personally judges to be the best for the helpee. The helper acts to optimize the well-being of the helpee. In the talent show example, you believe that your friend will utterly embarrass himself if he performs. With this belief in mind, you become anxious and feel a strong urge to help your friend by convincing him not to perform. This action of stopping your friend from performing is an example of paternalistic helping as it opposes their desire to perform. You stopping your friend is driven by a sense of worry that stems from an opinion formed by applying your unique perspective to your friend's situation. On the surface, paternalistic seems deeply related to the ability to reason about other's beliefs, but it is arguably more complicated. In the famous false belief task (Wimmer & Perner, 1983), the child only needs to select one belief to predict other's actions. In paternalistic helping, one not only needs to predict action with others' beliefs but also evaluate the actions with their own belief.

Based on paternalistic empathy, I will introduce our formulation of paternalistic coordination of mind, and a communication model on top of it. The model will also be compared against an individual perception model in a simulation experiment.

3.3 Individual Perception Model

We use a POMDP to model an intelligent agent with individual perception. As introduced in Chapter 2, a POMDP provides a generic formulation of how an agent takes rational actions in an uncertain environment with only limited observations. POMDP formulates the agent and the environment as the state space S, action space A, observation space omega, transition function $P(s'|s,a)$, reward function $R(s,a,s')$, and observation function $P(o|s',a)$. With these components, POMDP models the agent-environment interaction as two processes.

The first process is how an agent updates its belief of the world state s with observations o from the environment using Bayesian inference. A belief is defined as a probabilistic

distribution over the state space, the set of possible states $P(s)$. It serves as a prior before a new observation is received. The agent updates its belief when it gets an observation o after taking action a. When the agent takes an action a while in the world state s, the state transition to the next state s' with probability $P(s'|s,a)$ and the likelihood of observing a observation o from s' is $P(o|s',a)$. Therefore, by Bayes' rule, the agent's updated belief will be the posterior

$$P(s'|s,a,o) = \frac{P(s'|s,a)P(o|s',a)P(s)}{\sum_{s'} P(s'|s,a)P(o|s',a)P(s)} \tag{3.1}$$

This belief update only involves individual perception because the agent treats the observation only as being generated by its own interaction with the environment. It does not view it as a referent of any communicative intention as in (Grice, 1975), even with the presence of a second agent.

The second process is the decision model of how an agent takes rational actions based on its belief. Given an agent's belief, it can simulate the outcomes of taking an action. The outcomes include the states to visit and observations to receive in the future steps, how much more knowledge it can gain from the environment, and most importantly, how much reward can be cumulatively received. For the reward, the expected utility can be formulated as

$$Q(b,a) = \sum_{s'} P(s'|s,a)P(a)\left[R(s,a,s') + \gamma Q(SE(b,a,o),a)P(o|s',a)\right], \tag{3.2}$$

where SE is the state estimator in Eq. (2.12).

The process of deriving the optimal policy from a certain belief is introduced in Chapter 2 as solving a POMDP. In this study, we use the point-based value iteration (PBVI) algorithm (Pineau et al., 2003) as the POMDP solver. With a POMDP solver, we can obtain a probability distribution of policies

$$P(a|b) \propto \exp\{\alpha Q(b,a)\} \tag{3.3}$$

3.3.1 Joint Perception Model: Paternalistic Communication

In communication, two minds are involved. Our model of pointing adopts the philosophy that communication is cooperative, and in a paternalistic manner. The sender takes on the role of helper, while the receiver is the one who receives the help. Specifically, the sender uses theory of mind to estimate the receiver's belief, takes advantage of the receiver's belief to predict the receiver's actions, and then uses the sender's own belief to evaluate the receiver's actions. To reflect the paternalistic coordination of two beliefs in communication, we augment the formulation of single-agent decision-making with two subscripts: H representing the helper and P representing the player.

3.3.1.1 Policy prediction and policy evaluation

First, the sender predicts the receiver's actions with the receiver's belief. Here to simply, we assume that the sender knows the receiver's belief b_P. By augmenting Eq. (3.3), the receiver's policy can be predicted as

$$P(a|b_P) \propto \exp \alpha Q(b_P, a) \tag{3.4}$$

Here the notation $P(a|b_P)$ represents that the policy a is derived from belief b_P. If the receiver is completely rational, his policy can be predicted by

$$a = \arg\max_a Q(b_P, a) \tag{3.5}$$

Second, the sender evaluates the receiver's actions with her own belief. the subscript b_H represents that the expected utility is evaluated from the belief b_H.

$$Q(b_H, a) = \sum_{s'} P(s'|s, a) P_H(s) \left[R(s, a, s') + \gamma Q(b_H, a) P(o|s', a)\right] \tag{3.6}$$

Comparing the two equations, depending on the difference between b_H and b_P, the utility functions of the signaler and the receiver are consistent or different. This is consistent with

our discussion on paternalistic helping: the parent's evaluation of the outcomes may be consistent or different with their children's.

3.3.1.2 Smithian value of information

In the field of AI, information value theory enables an agent to choose what information to acquire. It assumes that, prior to selecting an instrumental action which can affect the physical world, the agent can acquire a piece of information. This action of seeking more information affects only the agent's belief state, not the external physical state. "The value of any particular observation must derive from the potential to affect the agent's eventual physical action; and this potential can be estimated directly from the decision model itself." (Russell, Norvig, & Davis, 2010). The definition of the value of a certain piece of information e_j is

$$VPI_e(e_j) = max_a EU(a|e, e_j) - max_a EU(a|e). \tag{3.7}$$

Here e can represent all the information collected by the agent prior to e_j, while the collection of e and e_j is all the information collected by the agent after acquiring the information e_j.

Acquiring a piece of information is similar to communication. The effect of communication is to change the receiver's belief. Therefore, to evaluate the effect of pointing, we can borrow the idea of evaluating the change of utility due to the change of the belief. We need to define the utility of a belief, especially the receiver's belief before and after the pointing.

If we have access to the belief of an agent, we can predict its policy as there is a mapping between belief and optimal policy. Therefore, the utility of a belief can be defined as the utility of the optimal policy derived from this belief.

$$U(b) = max_a Q(b, a). \tag{3.8}$$

In paternalistic helping, the utility of the receiver's policy is evaluated from the sender's

perspective. We can call this the paternalistic utility of belief, represented by $U_{b_H}(b_P)$.

$$U_{b_H}(b_P) = Q(b_H, \arg\max_a Q(b_P, a)). \tag{3.9}$$

With the definition of paternalistic utility of belief, we can define the Smithian value of information of a pointing signal as

$$SVI(u) = U_{b_H}(b_P^u) - U_{b_H}(b_P), \tag{3.10}$$

where b_P^u is the receiver's belief after receiving the communicative signal u.

The Smithian value of information is different from the value of information in information value theory. First, the Smithian value of information is more like the value of a specific information. The value of information is always non-negative, while the Smithian value of information can be positive, zero, or negative. If a communicative signal drives the receiver to a belief that values lower from the sender's perspective, then the Smithian value of information of this signal is negative. Second, the value of information is from a single-agent perspective, while the Smithian value of information involves the paternalistic coordination of the mind.

If we use belief-action value function Q to represent the utility of a belief-action pair, the Smithian value of information can be represented as $SVI(u) = max_a Q(b_P(u), a) - max_a Q(b_P, a)$. If the value of evidence is directly adopted, the utilities are all evaluated with the receiver's belief without the paternalistic coordination of mind. This will lead to a bad news paradox: If the goal of the communication is to improve the receiver's expected utility from the receiver's perspective, then bad news should never be told. That is because the receiver's original policy is already the optimal policy that is already high. Providing any bad news as a piece of evidence will drop the receiver's expected utility. This is obviously not true as we need to tell people bad news or correct their false cognition about the world.

The bad news paradox can be illustrated in the context of a hunting example: if a young hunter believes a prey is in front of him, he will have a high expected utility in shooting.

However, based on an experienced hunter's knowledge this is a bad action because she knows the prey is not nearby. If she uses value of evidence, she should not tell this piece of evidence at all because it would decrease the young hunter's expected utility. He would no longer expect to kill the prey, which is a huge drop of utility evaluated from the young hunter's perspective. But if the experienced hunter uses our paternalistic evaluation of beliefs, this evidence carries positive value because from the experienced hunter's perspective it prevents the young hunter from shooting and wasting an arrow.

3.3.1.3 Pointing as a rational action

With the utility of pointing clearly defined as SVI, we can treat pointing as a special type of utterance and model its use and interpretation using RSA. We can write down how a sender generates pointing u:

$$P_H(u|b_H) \propto \exp\{\alpha(SVI(u) - c(u))\}, \tag{3.11}$$

where $c(u)$ is the cost of sender the signal. For simplicity of the model, we can set $c(u) = 0$ as the cost of pointing in real world is small.

For the receiver, the Bayesian inference needs the likelihood $P_H(u|s)$. Here we assume that the sender knows the actual state of the world, so her belief b_H can be reduced to each state in the inference process. Then the likelihood $P_H(u|s) = P_H(u|b_H)$. With a prior b_P, the receiver's belief after receiving the pointing signal is

$$b_P^u(s) \propto b_P(s)P_H(u|s) \tag{3.12}$$

We can see the recursive nature of our model through the two equations above. In the sender's equation, the sender needs to estimate the receiver's belief after receiving the signal. In the receivers' equation, the receiver needs to estimate the sender's likelihood of sending out the signal. Therefore, we need a starting point to enter the recursion.

In this study, we use the context of whether one additional observation is useful as the entry point to the recursion. Similar to the literal receiver in RSA, our literal receiver

changes his mind by taking the pointed observation as a second observation. Then the signaler can start the recursion by comparing the two utterances: pointing and silence and their corresponding belief.

3.4 Simulation Task: the Guided Wumpus Hunt

3.4.1 The Wumpus World

To highlight the strengths of our model, we pick an augmented version of the classic AI problem: the Wumpus world (Russell et al., 2010) as a single-agent partially observable task as a baseline.

The Wumpus world is a game where a hunter navigates a cave to find a gold bar. The kill a terrible beast called the Wumpus lurks somewhere in the cave, waiting to eat anyone who enters its room in the cave. The hunter does not know where the Wumpus is but can infer it from its stench emitted to the adjacent rooms. There are also pits in the cave which will trap the hunter. They cause a breeze in their adjacent rooms. The hunter navigates the rooms in the cave to look for the stench, hence the Wumpus, while avoiding the pits. When the hunter is confident of the location of the Wumpus, he can shoot his only arrow to an adjacent room. If the Wumpus is in the targeted room, it will be killed. When the hunter enters the room with the gold bar, a reward will be granted to the hunter. The Wumpus world is not a difficult task for humans or computer programs but provides a good simulation for an agent navigating in a partially observable world. Formulating the game with a POMDP, the components are:

State space: Possible maps containing the hunter's location and direction, the location of the monster, the location of the gold bar, and whether each tile is a pit.

Action space: Move forward, Turn left, Turn right, Shoot

Observation space: None, Stench $\times$ None, Breeze $\times$ None, Glitter

Transition function: When the hunter moves or turns, he will deterministically move to intended location or turn to the intended direction. When the hunter moves to a room with a pit, the gold bar, or the Wumpus, the game ends. When the hunter hits the Wumpus, the Wumpus will be removed from the map. When the hunter misses the Wumpus, the state will not change.

Reward function: When the hunter moves or turns, he will pay a small navigation cost. When the hunter moves to a room with a pit or the Wumpus, he will get a large penalty. When the hunter moves to a room with the gold bar, he will get a large reward. When the hunter misses the Wumpus, he will get a small penalty.

Observation function: When the hunter enters the room next to the Wumpus, he can smell the stench. When the hunter enters the room next to a pit, he can feel the breeze. When the hunter enters the room with the gold bar, he can see a glitter.

3.4.2 The Guided Wumpus Hunt

The original game of the Wumpus world provides an environment for problem solving and decision making in partially observable environments. However, it is not very suitable for studying communication. It only has one single agent, and the observation perception is deterministic. Therefore, we made some changes to the game, and call it the guided Wumpus hunt.

In the guided Wumpus hunt game, a hunter navigates through a set of tiles to shoot a stationary monster, the Wumpus, without knowing its exact location. First, we remove the gold bar and the pits to simplify the state space. Instead of collecting the gold bar to receive the reward, now the hunter receives the reward if he kills the monster. We also reduce the dimension of the original Wumpus world from a square larger than 4×4 to a 3×3 triangle. The Wumpus can only be in the three tiles that are not or not adjacent to the origin, the starting point of the hunter. Here, although the state space is smaller

than the classic Wumpus world, the inference is in fact computationally more expensive as the POMDP solver is recursively called by RSA, and the stochastic observations to be introduced.

Secondly, we made the observations to the hunter stochastic. If the Wumpus is in an adjacent tile, a stench will be smelled with a probability p. The stochasticity supports the individual perception baseline. For a deterministic observation whose likelihood is either 0 or 1, a second observation will not help infer the state that generates the observation. However, for a stochastic observation, a second observation can help the inference better than the same but one observation.

Thirdly and most importantly, we add a second agent, the guide to help the hunter play the game. To the guide, the physical environment is fully observable. She can see the location of the hunter and the Wumpus. Her help to the hunter is through communication, but her communication to the hunter is minimal. She can only decide whether to point to a stench after the hunter observes it, without specifying whether the observation is accurate, what she hopes the hunter to do with that stench. We also assume that the helper knows the mind of the hunter. She can assume that the hunter always starts from a flat prior for the location of the Wumpus and has access to all the observations received from the physical world by the hunter. Then she can simulate the receiver's mind deterministically. If we want to break the assumption that the hunter starts from a flat prior, the helper will no longer have access to the hunter's belief. She can infer the hunter's belief by observing his actions and trajectories. The whole process will be a POMDP for the signaler, which we will further discuss in Chapter 6.

3.5 Simulation Experiment

3.5.1 Conditions

The setting of the guided Wumpus hunt game is directly inspired by the "broken stick" example, to capture the overloaded and indirect nature of pointing. We can use it to test our joint perception model against two baselines: the single-agent baseline and the double observation baseline.

We use the POMDP-based agent as the single-agent baseline model of a hunter who individually perceives the observation. In this condition, the hunter, as a receiver, will completely ignore the pointing signal sent by the helper as the sender. We use paternalistic pointing as the model of a hunter who pragmatically perceives the observation pointed to by the guide. The guide will use the Smithian pointing model to generate signals. Since the purpose of pointing is to help the receiver, the helper should send out signals that can increase the hunter's reward. Therefore, we predict that hunters who use the Smithian pointing model will outperform those who use POMDPs.

It is possible that this predicted improvement in performance is simply caused by the amount of information provided by pointing but not the pragmatic inference process: In the single-agent baseline, the hunter only receives one observation from the physical world; In the Smithian pointing condition, the hunter receives not only the one observation from the physical world, but also an additional communicative signal from the sender.

To test this possibility that an extra observation causes the reward increase, we add a third condition: a single POMDP-agent controlling the amount of information received. In this condition, the hunter uses POMDPs to individually perceive observations, but when receiving a pointing signal, he receives an additional observation to which the pointing is directed. We call this condition POMDPs with "double observations".

Among the Smithian pointing model and the two individual perception baselines: one

with single and one with double observations, we predict that joint perception model performs different and better than both individual perception models. Another variable is the hunter's difficulty of collecting more information in the environment. It is manipulated by varying the hunter's moving cost in the environment from -1, -3, -5, -7, and -9. A moving cost of -1 represents a scenario where collecting more information is easy because the hunter can collect more observations at a lower cost; A scenario with a moving cost of -9 is difficult for the hunter to collect information. We predict an improvement in performance consistent over all moving cost conditions.

3.5.1.1 Map

The map of the guided Wumpus hunt game is shown in Fig. 3.1. The hunter is an agent navigating the environment with a POMDP solver. Components of the POMDP framework are introduced below.

State Space: The Wumpus can show up in one of the three possible locations: $(0, 2)$, $(1, 1)$, and $(2, 0)$. Hunters will always start from (0, 0) and explore 3 possible locations for collecting observations: (0, 0), (0, 1), and (1, 0). Therefore, a tuple (L_P, L_W) can be used to represent all possible states, where $L_P \in \{(0, 0), (0, 1), (1, 0)\}$ represents the location of the player himself while $L_W \in \{(0, 2), (1, 1), (2, 0)\}$ represents the location of the Wumpus.

Action Space: The hunter can choose from 4 possible actions: move vertically, move horizontally, shoot to the upper tile, shoot to the right tile, which are the only possible directions given the map.

Transition Function: The outcomes of the hunter's actions are deterministic. At each step, he can decide to move or shoot. With moving, he will always move one tile in the direction of the action. If the hunter moves outside of the map, he will return to (0, 0). If the hunter chooses to shoot, an arrow will be shot into the adjacent tile to the direction of the shot. If the Wumpus is in the target tile, it is a hit; otherwise, it is a miss. The game

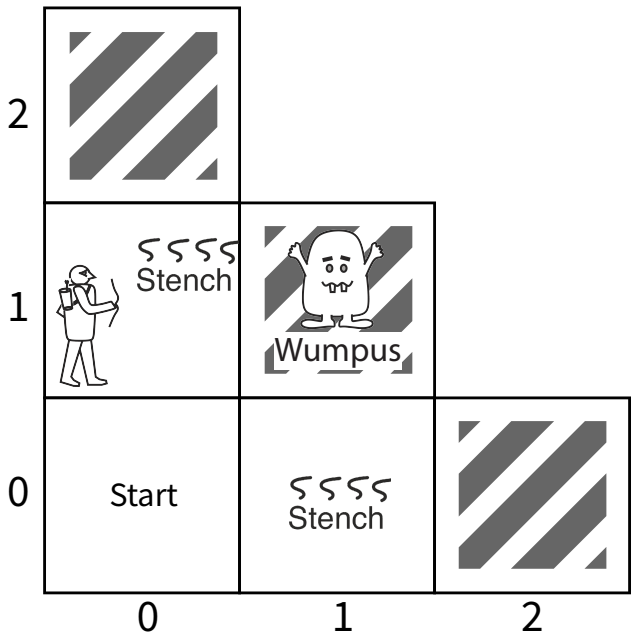

Figure 3.1: **The environment of the guided Wumpus hunting game.** Wumpus can only show up in one of three shaded tiles. Starting from (0, 0), the hunter tries to infer Wumpus's location from the stench and shoot it.

ends after one shooting action.

No matter what action is taken by the hunter, only his location L_P changes. The location of the Wumpus L_W will not change.

Reward Function: If the hunter moves one step, there is an action cost, which is manipulated from -1 to -9. If the hunter shoots the arrow, he will get a reward of 100 for hitting the Wumpus, or -100 for missing the Wumpus.

Observation Space and Observation Function: There are only two possible observations: a stench or nothing. The observation function is stochastic. If the hunter is in a tile adjacent to the Wumpus ($\|L_P - L_W\|_1 = 1$), the hunter's probability of observing a stench in that tile is 0.85 while the probability of observing nothing is 0.15. If the hunter is in a tile not adjacent to the Wumpus ($\|L_P - L_W\|_1 > 1$), the probability of observing a stench is 0.15 while the probability of observing nothing is 0.85. We do not treat communication as an observation in the observation space as it is modeled as a communicative act, instead of an observation generated from the physical world.

3.5.2 Result

The average reward across trials for each model under various moving cost is depicted in Fig. 3.2. Overall, the proposed Smithian pointing model achieves better performance compared to the classic POMDP model or the POMDP with double observations. The main effect of model type is significant ($F(2, 1485) = 9.602$, $p < 0.001$), and the main effect of moving cost is also significant ($F(4, 1485) = 19.535$, $p < 0.001$). However, the interaction between models and moving costs is not significant ($F(8, 1485) = 1.035$, $p = 0.407$). A post-hoc test with bonferroni correction shows that the Smithian pointing model achieves higher performance than "Double observation" condition ($F(1, 998) = 8.875$, $p = 0.009$). As hypothesized, our experiment also shows that the advantage stemmed from the pragmatic inference of pointing disappears when the task is too hard/easy for the guide to help.

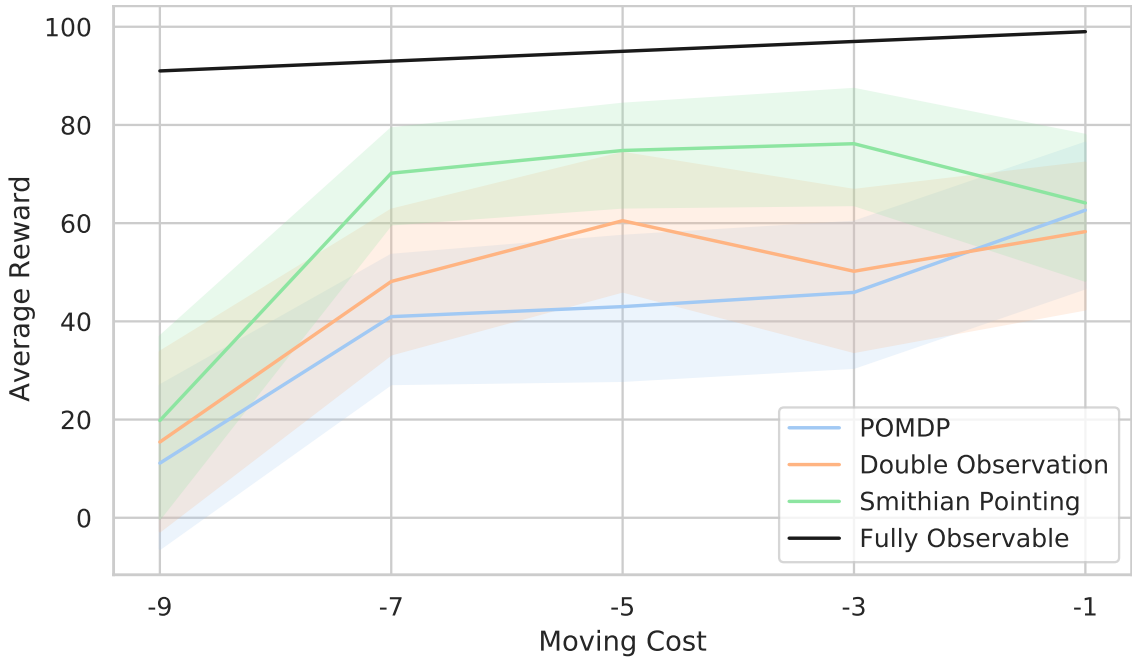

Figure 3.2: **Experimental results.** Black line denotes the ideal performance upper bound if the game was fully observable. Shaded areas represent 95% bootstrap confidence interval.

Specifically, when the moving cost is -1 or -9, the effect of model type is not significant ($F(2, 297) = 0.163$, $p = 0.850$; $F(2, 297) = 0.228$, $p = 0.796$). Taken together, these results demonstrate that pointing is relevant only when the signaler could offer help. Our computational model captures this relevance (Sperber & Wilson, 1986) and highlights how joint perception can be more powerful than individual perception of the same observation, demonstrated by the improved hunting performance.

3.6 Discussion

In this paper, we devise a computational model for pointing by defining the SVI, applying it to RSA to define the utility of pointing. In an example task, our pointing model shows a significant performance improvement compared to a single-agent POMDP or a single-agent POMDP with "double observation". This improvement indicates that the advantage of

pointing does not come from providing a new observation for individual perception. Instead, it comes from the pragmatic inference of how the jointly perceived observation is relevant to the task. Supporting the argument, our experiment also shows that the advantage of the Smithian pointing model works the best only when the receiver is in a position to be helped.

Our results suggest that paternalistic coordination of beliefs is necessary for modeling pointing. Seemingly simple, pointing requires the intelligent social capacity of mind coordination and altruistic motivation of helping. As a type of sophisticated social cognition, pointing involves identifying receiver's belief to predict action, evaluating actions with signaler's own belief, generating pointing to help change receiver's mind, and interpreting pointing with the assumption that the signaler is trying to help. The possible lack of some of these components, especially the cooperative motivation in pointing generation and interpretation, may explain the rarity of pointing in wild non-human primates. Indeed, the ability of generating and interpreting pointing is a milestone in human unique communication (Tomasello, 2010).

Due to the complexity of recursive reasoning of RSA, POMDP is solved multiple times in the recursion, which is computationally expensive and limits the current experiment setting to have a small state space. In follow up studies, we can use a faster POMDP solver or approximate the social recursion using computationally cheaper approaches (Kaelbling & Lozano-Pérez, 2013). In this way, we can expand our model to more complex tasks with larger state space and more observations. As many insights of this paper directly come from studies of child development, we hope this work will foster further interdisciplinary studies between developmental psychology and AI.

CHAPTER 4

A Relevance-based Communication Model for Human Overloaded Communication

Let's recall the hunting example from Chapter 2: Two hunters hunting in a forest. The young hunter sees a broken branch on the ground. Assuming that the branch was broken by the wind, he gets ready to continue his search. At this moment, his partner, the experienced hunter points the broken branch to him. The young hunter suddenly realizes that the branch was broken by their prey. He holds his breath and prepares to hunt.

The pointing signal sent by the experienced hunter, as a representative of human communication, is highly overloaded. The overloaded nature of human pointing is reflected in both referents and meanings. Pointing to a broken branch may refer to any object that aligns with the extension of the finger. Given that we know the pointing gesture refers to the branch, it may refer to one of multiple different features of the branch: its color, its shape, the fact that it is broken, or the fact that it is on the ground. This pointing signal to the broken branch may have so many different possible meanings without specifying the context: It can mean "We need that branch to start a fire"; It can mean "The branch is good material for a walking stick"; It can also mean "The branch was broken by the prey".

However, with proper context of hunting, the young hunter can pick the meaning "The branch was broken by the prey" accurately because it is the most relevant. If the context is that they are camping in the wilderness, then "We need that branch to start a fire" will be more relevant; If the context is that they are hiking together in the mountains, then "The branch is good material for a walking stick" will be more relevant. Different meanings are

interpreted when communicators engage in different joint tasks. The key is whether each meaning is relevant to the joint task.

Pointing is built on the mutually acknowledged assumption that human communication must be relevant (Sperber & Wilson, 1986). The sender has the obligation to send out relevant information. Infants as young as 12 months prefer to share information only when it matches their partner's current goal: when infants watch an adult misplace an object and search for it, they point to the exact object more often than to other objects not needed by the adult (Liszkowski et al., 2006). Meanwhile, the receiver has to also resonate with the relevance of what is being pointed to. Infants point significantly more to responsive adults than to ignorant ones; when the adult expresses disinterest, children no longer repeat the gesture (Carpenter & Liebal, 2011).

Mathematically, this can be formulated by evaluating the relevance of each possible meaning $m \in M$, and the most relevant meaning is the signaler's intended meaning m^*,

$$m^* = \arg\max_m Relevance(m, task), m \in M \tag{4.1}$$

4.1 Relevance in Linguistics Context

Now that the importance of relevance in human communication has been acknowledged, the real challenge is how to define it. As we introduced in Chapter 1, classic information theory does not focus on relevance in its models of communication. They focus on retrieving the accurate signal from a noisy channel and then interpret it with a codebook (MacKay, 2003). In the field of machine learning, one mainstream approach is to define relevance as the associations between variables. It is usually trained in a data-driven fashion from a large dataset. As an exemplar of this model family, deep learning-based models in natural language processing and computer vision analyze the mapping from syntax or visual stimuli to meaning (Collobert et al., 2011; Vaswani et al., 2017; Devlin et al., 2018; Dosovitskiy et al., 2020). Like models in information theory, these models do not focus on overloaded

signals. They rely on one-to-one mappings between signal and meaning, which is not how people understand overloaded human communications like pointing.

These association models of relevance also fail to capture that human communication is causally transparent (Pearl & Mackenzie, 2018). When interpreting pointing, the receiver should not rely on a massive training on interpretation, but an understanding of the underlying model of how pointing is generated. To understand the pointing act, the young hunter must ask the key questions: "Why was the pointing signal sent? What would I have done if I did not receive the point?"

The concept of relevance plays a crucial role in communication, enabling human pointing to be intuitively generated and understood in daily life. Grice (1975) introduced the maxim of relevance as a part of cooperative communication maxims, asserting that speakers must provide information that is relevant to the current conversation if they wish for listeners to correctly understand their meaning. Similarly, listeners can successfully understand the intended meaning when they assume the speaker is adhering to this maxim. In Grice's definition, a sentence is relevant if its meaning continues the dialogue. For example, one asks where the scissors are and receives the response "I cut paper in my room". Even though nothing about where the scissors are appears in the response, it is assumed that the person who responded has the responsibility to make the response relevant, which means that the response continues the discussion about where the scissors are. Therefore, it implied that I used the scissors to cut paper in my room so the scissors are in my room. In the Gricean theory, relevance is most obviously context-dependent, thus important in understanding contextual reasoning (Benotti & Blackburn, 2011).

Adding to Gricean definition of relevance, Wilson and Sperber (2006) proposed that relevance can be assessed in terms of cognitive effect and processing effort. In their definition of relevance, relevant information helps the listener by creating a worthwhile difference in their understanding of the world. Take the scissors as example, the response is relevant not only because it continues the discussion to answer the question, but also it helps who asked

the question have a better understanding of where the scissors are.

These insightful definitions of relevance are deeply planted in discussions of language. However, to study the relevance in visual communication like pointing, we need a definition of relevance that is based on agency. It should include the agent's coordination of mind and intelligent planning to interact with the environment. Therefore, we devise a causal model based on agency and utility calculus. More specifically, the causal process of signal generation and interpretation is modeled by how the agents' actions, both instrumental and communicative, are generated and driven by their mental states.

We start from the causal interpretation of relevance by defining relevance as to what degree a receiver's well being can be improved by a signaler sharing her belief. In communication, signalers tend to be altruistic (Tomasello, 2010). There is no point in a signaler telling a receiver information that does not make a difference to the receiver's well-being. Relevance must serve as an intervention to change the receiver's well being.

Under this assumption, the causal model of agency and utility theory are important for defining relevance. A sender must know their receiver's mental state and predict their actions and the consequences of those actions, as these are key to evaluating a receiver's well being.

4.2 Model

In Chapter 3, we introduced the concept of Smithian value of information. Different from value of information in information value theory, the Smithian value of information is a statistic adopted by the sender to select what signal to send. The effect of each communicative signal is to change the receiver's belief. Therefore, the Smithian value of information is defined as the sender's evaluation of the change of utility due to the change of the belief.

The sender's evaluation of the utility of a belief is what we call paternalistic evaluation. Paternalistic evaluation involves two consecutive processes: **policy prediction** and **policy evaluation**. If the sender has access to the belief of the receiver, the sender can predict the

receiver's policy as there is a mapping between belief and optimal policy. In paternalistic helping, the utility of the receiver's policy is evaluated from the sender's perspective. This utility is the paternalistic utility of belief, represented by $U_{b_H}(b_P)$. As in Eq. (3.9)

$$U_{b_H}(b_P) = Q(b_H, \arg\max_a Q(b_P, a)). \tag{4.2}$$

The Smithian value of information of a communicative signal is defined in as Eq. (3.10),

$$SVI(u) = U_{b_H}(b_P^u) - U_{b_H}(b_P), \tag{4.3}$$

where b_P^u is the receiver's belief after receiving the communicative signal.

SVI serves as the utility of a communicative signal, with which a sender generates pointing u:

$$P_H(u|b_H) \propto \exp\{\alpha(SVI(u) - c(u))\}, \tag{4.4}$$

where $c(u)$ is the cost of sender the signal. The receiver, on the other hand, can use $P_H(u|s)$ as the likelihood in Bayesian inference. With the recursive nature of our model, we need a starting point to enter the recursion. In Chapter 3, we use the context of whether one additional observation is useful as the entry point to the recursion. The literal receiver changes his mind by taking the pointed observation as a second observation. Then the signaler can start the recursion by comparing the two utterances: pointing and silence and their corresponding belief. However, this entry point does not solve the problem for all communicative signals. Many signals do not have its literal meaning, for example, a shout "Ahh!" or an eye gaze. However, we can also make rich and efficient inference from these signals. To accommodate for these signals, we define relevance as how much the signal can help its receiver, regardless of its literal meaning or even whether it has a literal meaning or not. We will explain our formulation with the hunting example.

4.2.1 Definition of relevance

Relevance is evaluated in multiple steps. First, the sender predicts the receiver's actions based on the receiver's belief. In the hunting example, the young hunter believes that the

wind broke the branch. Knowing this, the experienced hunter predicts that he will walk away. We can write down this action prediction as

$$a_P = \arg\max_a Q(b_P, a) \tag{4.5}$$

Next, the sender evaluates the predicted action with her own belief. The experienced hunter knows that the prey is around, so she knows that the hunt will be ruined if the young hunter walks away. In this case, the sender's evaluation of the receiver's action $Q(b_H, a_P)$ is low. Here, we can see the discrepancy of the evaluation of the same action based on different beliefs; a_P is the best choice of actions to the receiver, while it may be undesirable to the signaler.

The sender can help the receiver most if the receiver knows what the sender knows. In this case, receiving the signaler's mind, the rational receiver will take the action that maximizes $Q(b_H, a)$, which will improve his well-being as evaluated by the sender to the most. In the hunting example, if the young hunter knows that the prey is around, he will absolutely stay silent and prepare to hunt. To the experienced hunter, this action is much better than the young hunter's original plan. In this case, the sender's belief is relevant.

Formally, we define the relevance of the sender's mind to the receiver as the gap between the best outcome a receiver can obtain and the outcome that the receiver is going to get without any interaction with the sender, both evaluated with the sender's mind.

$$Rel(b_P|b_H) = max_a Q(b_H, a) - Q(b_H, \arg\max_a Q(b_P, a)). \tag{4.6}$$

Sharing information must make a difference. More specifically, sharing information must improve the receiver's well-being. This is the key to interpret many signals, even without a literal meaning. For example, imagine an adult point to a dress to a child. One possible interpretation of this pointing is the dress is made from polyester. However, this interpretation is not relevant. It makes no difference to the child's well-being because knowing a dress is made from polyester is helpful to the child. It is more likely to have the meaning do you

want to buy that dress. On the other hand, if the receiver of this pointing is a designer or a child interested in designing, the "the dress is made from polyester" interpretation may be relevant, as knowing the fabric may increase a designer's well-being.

In practice, the receiver may not recover the signaler's exact belief. Instead, he may interpret the signaler's belief as b_P^u. Formally, we define the utility of pointing as the utility change before and after communication, evaluated based on the sender's belief. For a pointing signal u, the utility is

$$Rel(u) = Q(b_H, \arg\max_a Q(b_P^u, a)) - Q(b_H, \arg\max_a Q(b_P, a)) \tag{4.7}$$

The relevance definition in Eq. (4.7) can be used as the entry-level utility in the recursion of RSA. With the utility of pointing clearly defined, we can model sending out each pointing signal as a rational action. Relevance is directly connected to the instrumental utility change caused by the signal. Therefore, it can be used as the utility in RSA for the inference of meanings.

4.2.2 Relevance as utility in communication

In this inference problem, since we define relevance based on two beliefs, both the sender and the receiver need to We first establish a belief system that takes into account not only the physical state of the world but also the beliefs held by the other agent. In our task, the player lacks knowledge of both the physical state and the helper's belief. Therefore, the player's belief b_P can be represented through a joint probability distribution of s and the helper's belief b_H. This state including both physical state and belief state is an interactive state $is_P = (s, b_H)$, as defined in interactive-POMDP (P. J. Gmytrasiewicz & Doshi, 2005). The receiver's belief is a joint probability distribution over the physical world state and the belief state of the sender.

$$P(IS_P) = P_{S,B_H}(S = s, B_H = b_H | b_P) \tag{4.8}$$

In our task, the helper knows the world s^* and the player's the belief b_P. The player's belief b_P is a flat prior on both the physical state and the helper's belief. However, since he knows that the helper has full knowledge of the physical state, the helper's belief and the physical the helper's belief reflects the true state. Therefore, his belief is $P(IS_P) = P_{S,B_H}(S = s, B_H = b_H|b_P) = P_S(s|b_P)I(b_H = \delta(s))$. His belief about the physical state is the marginal distribution $P_S(s|bP)$ and his belief about the sender's belief state is the marginal distribution $P_{B_H}(b_H|b_P)$.

Then the probability that the signaler takes the action of pointing u is

$$P(u|b_H, b_P) \propto \exp\{\alpha Rel(u)\}. \tag{4.9}$$

The receiver's interpretation of the pointing signal can be modeled with Bayesian update:

$$P(s, b_H|u, b_P) \propto P(s, b_H|b_P)P(u|b_H, b_P). \tag{4.10}$$

The state of the world can be inferred by

$$P(s|u, b_P) = \sum_{b_H} P(s, b_H|b_P). \tag{4.11}$$

If the sender knows the exact state of the world, then the Bayesian update reduces to

$$P(s|u, b_P) \propto P(s|b_P)P(u|s, b_P). \tag{4.12}$$

4.2.3 Relevance in the POMDP belief space

As we introduced in Chapter 2, a belief of an agent in a POMDP can be represented as a probability distribution $b \in \Delta S$ over the state space. In this section, we use the components in POMDP to understand the definition of relevance.

On the belief space ΔS, the value function of POMDP $V : \Delta S \to \mathbf{R}$ is the maximum expected future reward that could be obtained by an agent who owns this belief. The value function is convex on the belief space. Each policy π of the agent can be represented by

a vector $\alpha_\pi \in \mathbf{R}^{|S|}$, whose elements are the evaluation of the policy on each state. The evaluation of the policy based on a belief b can be represented as the inner product of the belief b and α_π.

$$Q(b, \pi) = b^T \alpha_\pi. \tag{4.13}$$

On the differentiable points of the value function V, the gradient ∇V represents the optimal policy π^* for this belief.

$$V(b) = Q(b, \pi^*) = max_\alpha b^T \alpha = b^T \Delta(b). \tag{4.14}$$

The relevance in our definition is

$$Rel(b_P|b_H) = V(b_H) - b_H^T \nabla V(b_P). \tag{4.15}$$

When the helper knows the true state of the world s, the relevance is $Rel(b_P|s) = V(b_H) - \nabla V(b_P)(s)$. For the relevance of a signal u, we only need to change the helper's belief b_H to the player's belief after receiving the signal b_P^u.

4.3 Simulation Experiment 1

We test the relevance model of pointing with an augmented version of the classic AI task, the Wumpus world (Russell et al., 2010), which is partially observable. In the Wumpus world, a hunter tries to kill a monster called Wumpus. However, he cannot see the location of the Wumpus and can only infer its location by his observation of the stench it emits. To simulate communication, we add another agent, the guide, who observes everything about the environment and the hunter. However, the only way she can communicate to the hunter is to point to an observation the hunter has already observed. The hunter needs to infer the meaning of the pointing and act accordingly. We call this game the *guided Wumpus hunting*. It is inspired by the hunting example, with highly sparse, overloaded, and indirect communication.

We conduct two simulation experiments with the *guided Wumpus hunting*. We start by testing our model of relevance with one observation. Of note, there is no uncertainty about which observation the point is referring to. The second experiment raises this challenge by adding another observation, incorporating overloadedness into the experiment. In both experiments, we compare the performance of agents with our relevance model and agents who use a single agent model as the baseline. In both models, the belief-action value function $Q(b, a)$ is calculated by a POMDP solver, the PERSEUS algorithm (Spaan & Vlassis, 2005). In addition, to distinguish the relevance model from enhancing individual perception, we add a control condition called "double observation." In this condition, the agent uses the single agent model, but he receives a second observation from the environment as if he observed the world twice.

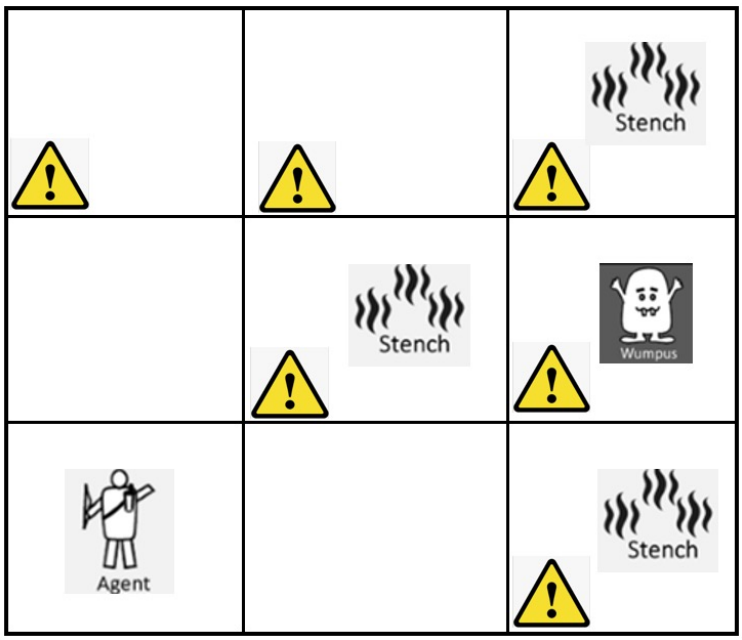

Figure 4.1: **Environment for Experiment 1.** The Wumpus may locate in one of the six tiles with the warning sign.

4.3.1 Task

In this simulation, we use the guided Wumpus hunt from Chapter 3.4, but expand it to a 3×3 square grid. The layout of the grid is shown in Fig. 4.1.

The hunter starts from the bottom left tile (0, 0). The Wumpus stays in one of the six locations that are not the start or the tiles adjacent to the start ($\{(0,2),(1,1),(1,2),(2,0),(2,1),(2,2)\}$, the tiles with a warning sign), and emits a stench to the nearby tiles. The hunter does not know the exact location of the Wumpus but can infer its location from the stenches. The hunter's observation is not 100% accurate but will be controlled as a variable.

The hunter starts from the bottom-left corner tile. He can move or shoot in four directions: up, down, left, and right. A moving action will move the hunter one tile in the selected direction. If the action moves the hunter outside of the map, the hunter will not move. A shooting action will shoot an arrow to the adjacent tile in the selected direction. The hunter can move unlimited steps in the map, but moving each step has an action cost of 5. The game ends when the hunter shoots or enters the tile of the Wumpus. If the hunter moves to the tile of the Wumpus, he gains -100. If he shoots and hits the Wumpus, he will gain a reward of 100. However, if he misses the shot, he will get -100. If the hunter moves to the same tile as the Wumpus, he will get -100 too.

The helper knows the location of the Wumpus and shares the reward with the hunter. She tries to help the hunter shoot the Wumpus by communicating with one communication signal. We can compare this to a shout "Ahh!".

The hunter cannot see the Wumpus, but he can infer the location of the Wumpus by observing its stench. There are two possible observations in the environment, stench or nothing. If the hunter is in a tile next to the Wumpus, he will have a high probability $p > 0.5$ of observing the stench. If the hunter is not next to the Wumpus, he will have a low probability $1 - p$ of observing the stench. The observation accuracy p is manipulated as an experiment condition. In the classic Wumpus world, $p = 1$. We add more stochasticity to

increase the task difficulty and the need for pointing. Although we only have one possible referent to point to, which is the stench, the pointing is still overloaded. It can mean that the Wumpus is in one of all possible tiles (see Fig. 4.2) or that the hunter should take one of all possible actions.

4.3.2 Map

The map of the guided Wumpus hunt game is shown in Fig. 4.1. The hunter is an agent navigating the environment with a POMDP solver. Components of the POMDP framework are introduced below.

State Space: The Wumpus can show up in one of the six possible locations: (0, 2), (1, 1), (1, 2), (2, 0), (2, 1), and (2, 2). Hunters will always start from (0, 0) and explore all 9 possible locations for collecting observations. Therefore, a tuple (L_P, L_W) can be used to represent all possible states, where $L_P \in \{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(2,0),(2,1),(2,2)\}$ represents the location of the player himself while $L_W \in \{(0,2),(1,1),(1,2),(2,0),(2,1),(2,2)\}$

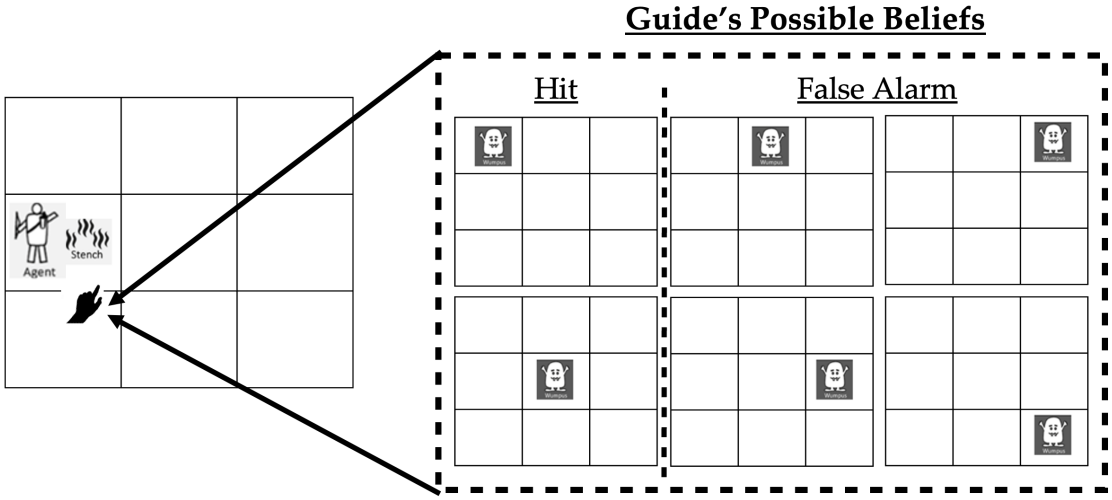

Figure 4.2: **Overloadedness of the pointing act in Experiment 1.** The pointing to the stench can mean an accurate observation or a false alarm, each leading to multiple possible world states.

represents the location of the Wumpus. When $L_P = L_W$, the state is the same as *endgame* in Chapter 3.

Action Space: The hunter can choose from 8 possible actions: {move up, move down, move left, move right, shoot up, shoot down, shoot left, shoot right}.

Transition Function: The outcomes of the hunter's actions are deterministic. At each step, he can decide to move or shoot. With moving, he will always move one tile in the direction of the action. If the hunter moves outside of the map, he will stay at his current location. If the hunter chooses to shoot, an arrow will be shot into the adjacent tile to the direction of the shot. If the Wumpus is in the target tile, it is a hit; otherwise, it is a miss. If the hunter shoots towards the wall, there is no adjacent tile to the direction of the shot, and it will count as a miss. The game ends after one shooting action or if the hunter moves to the same tile as the Wumpus. No matter what action is taken by the hunter, only his location L_P changes. The location of the Wumpus L_W will not change.

Reward Function: If the hunter moves one step, there is an action cost of -5. If the hunter shoots the arrow, he will get a reward of 100 for hitting the Wumpus, or -100 for missing the Wumpus. If the hunter moves to the same tile as the Wumpus, he will get a reward of -100.

Observation Space and Observation Function: There are only two possible observations: a stench or nothing. The observation function is stochastic. If the hunter is in a tile adjacent to the Wumpus ($\|L_P - L_W\|_1 = 1$), the hunter's probability of observing a stench in that tile is p while the probability of observing nothing is $1 - p$. If the hunter is in a tile not adjacent to the Wumpus ($\|L_P - L_W\|_1 > 1$), the probability of observing a stench is $1 - p$ while the probability of observing nothing is p. The value of p is manipulated from 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, and 1. We do not treat communication as an observation in the observation space as it is modeled as a communicative act, instead of an observation generated from the physical world.

4.3.3 Conditions

The experiment is a 3×7 design; there are three models and seven observation accuracies.

Similar to Chapter 3, we use two baselines: the single-agent baseline and the double observation baseline to test our relevance-based communication model. We use the POMDP-based agent as the single-agent baseline model of a hunter who individually perceives the observation. In this condition, the hunter will completely ignore the pointing signal sent by the helper.

We also use the double observation condition in Chapter 3 as a single-agent baseline. In this condition, the hunter uses POMDPs to individually perceive observations, but when receiving a pointing signal, he receives an additional observation to which the pointing is directed. This condition aims to control the amount of information received by the hunter.

We use relevance-based paternalistic pointing as the model of a hunter. The hunter pragmatically perceives the observation pointed to by the helper with the relevance model. The guide will use the relevance model to generate signals.

We also manipulated the accuracy of the hunter's observations by 7 levels: $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1\}$. For example, if the observation accuracy is 0.8, then the hunter has a 0.85 probability of observing the stench and a 0.15 probability of observing nothing when the Wumpus is nearby, and a 0.15 probability of observing the stench and a 0.85 probability of observing nothing when the Wumpus is not nearby. We expect that the relevance model can increase the received reward more when the observations are less accurate.

We run 100 game simulations for each condition and record the reward gained by the hunter. We predict that a) the hunters who use the relevance model will gain more rewards than the hunters who use the other two models, and b) as the observation accuracy decreases, the performance of the relevance model does not decrease because the power of our relevance model comes from ToM inference instead of observation accuracy.

4.3.4 Results

The average reward across trials for each model under various observation accuracies is depicted in Fig. 4.3. Overall, agents who use the proposed relevance model achieve a higher average reward than agents who use the single agent POMDP model or the double observation model. The main effect of model type is significant ($F(2, 2079) = 141.926$, $p < 0.001$), and the main effect of observation accuracy is also significant ($F(6, 2079) = 58.370$, $p < 0.001$). The interaction between models and observation accuracy is significant ($F(12, 2079) = 16.303$, $p < 0.001$). A post-hoc test with Bonferroni correction shows that agents who use the relevance model of pointing gain a higher reward than agents who use the double observation model ($F(1, 1398) = 109.882$, $p < 0.001$). Our results show the power of relevance-based pointing, especially when the observation accuracy is low. It is more effective in helping its receiver than providing more accurate observations to his individual attention.

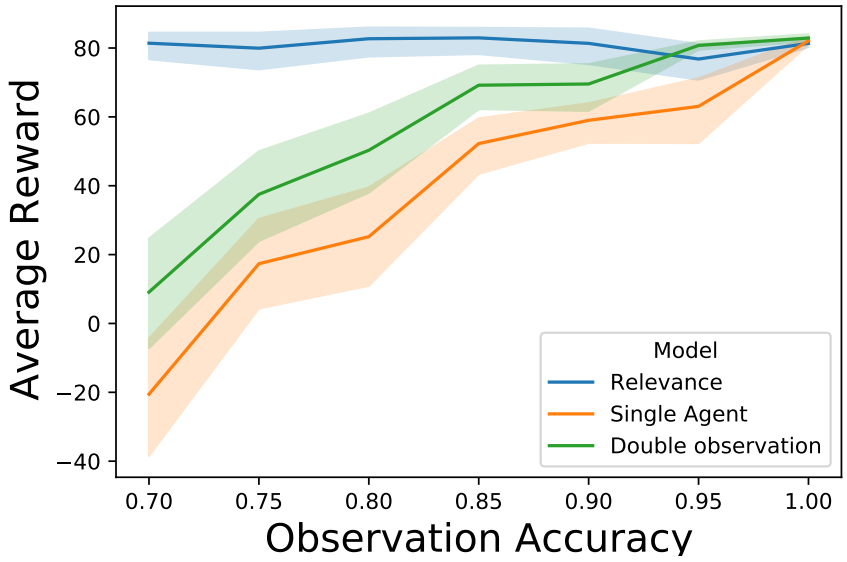

Figure 4.3: **Results of Experiment 1.** Shaded areas represent 95% bootstrap confidence interval.

4.4 Simulation Experiment 2

In Experiment 1, we have only one type of feature in the observation: the stench. Although the pointing signal may still have multiple interpretations as shown in Fig. 4.2, it has only one referent. This lacks the overloadedness discussed by Wittgenstein and Anscombe (1953/2001). To capture this overloadedness, we added another observation of glitter to the environment. Here, the glitter and stench coexist in the same grid. This setting offers a more complete evaluation of the Wumpus world.

4.4.1 Task

The setup of Experiment 2 is identical to Experiment 1 except for a few aspects. We reduced the size of the environment because including an additional observation exponentially

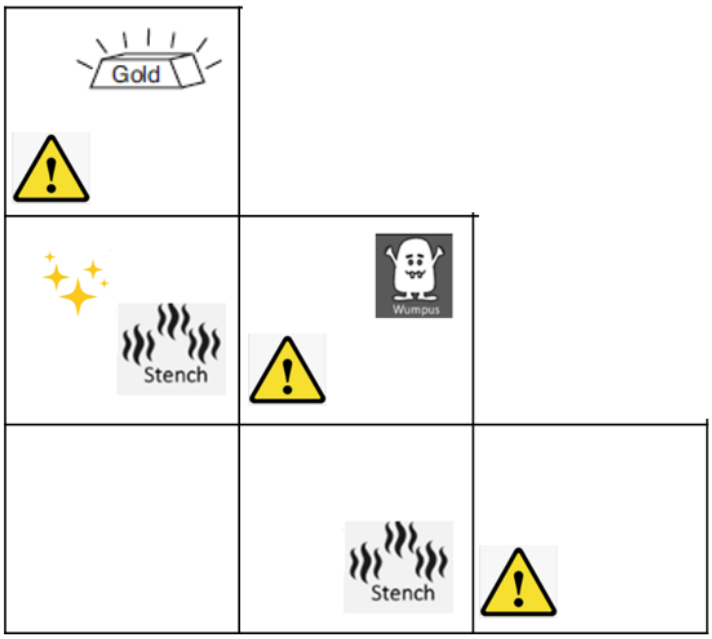

Figure 4.4: **Environment for Experiment 2.** The Gold bar and the Wumpus locate in two of the three signed tiles.

decreased the speed of our POMDP solver. The environment of Experiment 2 is shown in Fig. 4.4. A gold bar is added to the game as a source of the glitter. Picking up the gold bar gives the hunter a reward. The Wumpus and the gold bar are located in two different tiles of the three tiles with a warning sign. They do not move. They are invisible to the hunter but visible to the guide.

The hunter starts from the bottom-left corner tile. His moving and shooting actions have the same effect as in Experiment 1. In addition, the hunter has another action of picking up the gold bar. This action removes the gold bar if the hunter is in the same tile with the gold bar. Otherwise, it does not change the environment. The action of picking up the gold bar has a cost of 5 if he misses the gold bar. The hunter will gain a reward of 100 if he successfully picks up the gold bar.

The gold bar spreads glitter to its nearby tiles. The observations of glitter and stench of the Wumpus have the same probability model as the stench in Experiment 1. The observations of the stench and the glitter are independent, resulting in four possible observations in the environment. For example, if the hunter is in a tile that is adjacent to the Wumpus but not the gold, he may observe a) both glitter and stench with probability $p(1-p)$, b) single glitter with probability $(1-p)^2$, c) single stench with probability p^2, or d) nothing with probability $p(1-p)$.

4.4.2 Map

The map of the guided Wumpus hunt game is shown in Fig. 4.4. The hunter is an agent navigating the environment with a POMDP solver. Components of the POMDP framework are introduced below.

State Space: The Wumpus and the gold bar can show up in two of the three possible locations: (0, 2), (1, 1), and (2, 0). Hunters will always start from (0, 0) and explore all 6 possible locations for collecting observations. Therefore, a tuple (L_P, L_W, L_G) can be used to rep-

resent all possible states, where $L_P \in \{(0,0),(0,1),(0,2),(1,0),(1,1),(2,0)\}$ represents the location of the player himself while $(L_W, L_G) \in \{((0,2),(1,1)),((0,2),(2,0)),((1,1),(0,2)),((1,1),(2,0)),((2,0),(0,2)),((2,0),(1,1))\}$ represents the location of the gold bar and the Wumpus. The state space can be reduced because when $L_P = L_W$ is the same as endgame.

Action Space: The hunter can choose from 9 possible actions: {move up, move down, move left, move right, shoot up, shoot down, shoot left, shoot right, pickup}.

Transition Function: The outcomes of the hunter's actions are deterministic. At each step, he can decide to move or shoot. With moving, he will always move one tile in the direction of the action. If the hunter moves outside of the map, he will stay at his current location. If the hunter chooses to shoot, an arrow will be shot into the adjacent tile to the direction of the shot. If the Wumpus is in the target tile, it is a hit; otherwise, it is a miss. If the hunter shoots towards the wall, there is no adjacent tile to the direction of the shot, and it will count as a miss. The game ends after one shooting action or if the hunter moves to the same tile as the Wumpus. No matter what move or shoot action is taken by the hunter, only his location L_P changes. The location of the Wumpus L_W and the location of the gold bar L_G will not change. If the hunter is in the same tile as the gold bar and he takes the action pick up, this action will remove the gold bar from the map.

Reward Function: If the hunter moves one step, there is an action cost of -5. If the hunter shoots the arrow, he will get a reward of 100 for hitting the Wumpus, or -100 for missing the Wumpus. If the hunter moves to the same tile as the Wumpus, he will get a reward of -100. The action of picking up the gold bar has a cost of 5 if he misses the gold bar. The hunter will gain a reward of 100 if he successfully picks up the gold bar.

Observation Space and Observation Function: There are only four possible observations: a stench, a glitter, both the stench and glitter, or nothing. The observation function is stochastic and independent with regard to stench and glitter. If the hunter is in a tile adjacent to the Wumpus ($\|L_P - L_W\|_1 = 1$), the hunter's probability of observing a stench in that tile is p while the probability of not observing the stench is $1 - p$. If the hunter is in

a tile not adjacent to the Wumpus ($\|L_P - L_W\|_1 > 1$), the probability of observing a stench is $1 - p$ while the probability of not observing the stench is p. The same is true for gold bar and glitter. If the hunter is in a tile adjacent to the gold bar ($\|L_P - L_G\|_1 = 1$), the hunter's probability of observing a glitter in that tile is p while the probability of not observing the glitter is $1 - p$. If the hunter is in a tile not adjacent to the gold bar ($\|L_P - L_G\|_1 > 1$), the probability of observing a glitter is $1 - p$ while the probability of not observing the glitter is p. The value of p is manipulated from 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, and 1. We do not treat communication as an observation in the observation space as it is modeled as a communicative act, instead of an observation generated from the physical world.

4.4.3 Conditions

The design and conditions are the same as Experiment 1. We predict that there will be an advantage in rewards with the relevance model compared to a single agent model and double observation model.

4.4.4 Results

The average reward across trials for each model under various observation accuracies is depicted in Fig. 4.5. Similar to Experiment 1, agents who use the relevance model achieve a higher reward on average than agents who use the single-agent POMDP model or the double observation model. The main effect of model type is significant ($F(2, 2079) = 41.732$, $p < 0.001$), and the main effect of observation accuracy is also significant ($F(6, 2079) = 20.049$, $p < 0.001$). The interaction between models and observation accuracy is significant ($F(12, 2079) = 9.130$, $p < 0.001$). A post-hoc test with Bonferroni correction shows that agents who use the relevance model of pointing gain higher reward than agents who use the double observation model ($F(1, 1398) = 21.163$, $p < 0.001$). Our results are consistent with the results in Experiment 1. The relevance model of pointing still achieves high performance

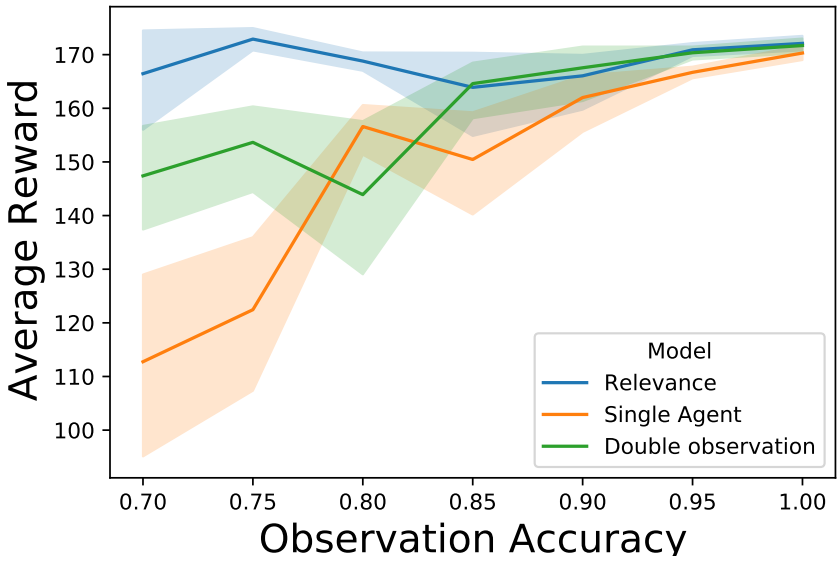

Figure 4.5: **Results for Experiment 2.** Shaded areas represent 95% bootstrap confidence interval.

in highly overloaded communication.

4.5 Discussion

Our relevance-based communication model was successful in capturing the essence of transparent communication. In both Experiment 1 and Experiment 2, the agents who used the relevance model for communication achieved better performance than agents who ignored the communication. In addition, agents who used the relevance model achieved better performance than agents with more precise individual perception by having two samples from the environment. The high performance of the relevance model was robust over all observation accuracies.

Our results showed that by leveraging ToM and utility theory, agents can achieve overloaded communication without a predefined codebook. In the experiments, the guide and the

hunter do not have a codebook that regulates the relevance between signals and meanings. Crucially, they never learned the relevance through massive training. They promptly calculate the relevance based on their context of the cooperative task. Our results also showed that the power of pointing comes from the ToM inference which supports the relevance calculation. The power of pointing does not come from enhancing individual perception, which is supported by the robust performance of the relevance model across all observation accuracies.

CHAPTER 5

Predict Human Communication with Relevance-based Model

5.1 Introduction

In a football game, the receiver catches the ball. Despite three opponent defenders rapidly approaching him, he accelerates without noticing them. Within that split second, you can only alert him about one of the opponents. How do you make that choice?

Living in a fast-paced modern world, we constantly face an overload of information while needing to make quick decisions of what to communicate. This necessity for rapid decision-making applies to contexts as diverse as emergency room operations, stock trading, and professional kitchens. In these domains, communication must be swift and effective, conveying substantial information efficiently.

But how do humans manage to distill complex information into concise, impromptu communicative signals? How do we determine what information to share? In this paper, we use pointing—a remarkably simple form of communication—as a case study to investigate how humans swiftly and effectively choose which information to communicate.

Pointing is overloaded, indirect, and sparse, and yet it is expected to be accurately interpreted in a mere instant. The context in which we are engaged imposes a critical constraint on how we select actions and consequently interpret pointing gestures. Assuming that humans are rational, it is imperative for them to perform pointing signals that facilitate achieving the highest utility in the most efficient manner possible. Given the spontaneous

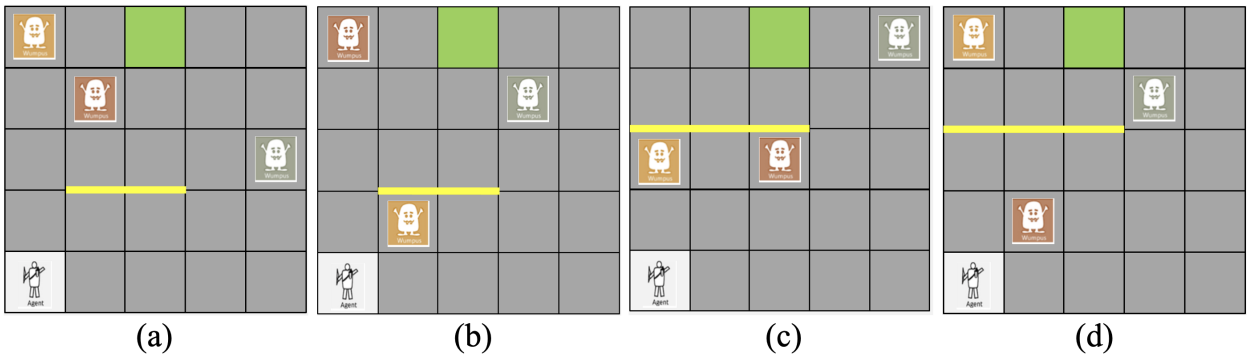

Figure 5.1: **4 examples of maps**: (a) Green not on any shortest path; (b) Orange and Green not on shortest paths; (c) Orange blocked by Red; (d) Red blocked by Green; Green high in relevance.

nature of contextual constraints, pointing becomes extremely flexible in meaning and adapts to the environment.

Recent studies in cognitive science have indirectly implied the concept of relevance. The RSA model (Frank & Goodman, 2012), for instance, is a linguistic model developed from reference games, in which one attempts to find a target among distractors based on a description. The RSA model assigns each utterance a utility based on how much it can direct the listener's belief towards the true target. Extending this reference game to decision-making, relevance emerges as an utterance's ability to increase the listener's reward. Studies on pedagogy reveal that individuals meticulously select their instructions by considering its impact on the listener's performance and belief (T. Sumers, Ho, Griffiths, & Hawkins, 2022; Ho et al., 2016). Another study shows the proficiency of robot assistants in providing the most helpful–even if somewhat distorted–observations to aid task-solving (Reddy et al., 2021). Our prior research demonstrates that basing signal generation and inference on the principle of helpfulness significantly improves agent performance, overcoming the challenges posed by the extremely overloaded signal (Jiang et al., 2022). To further our understanding of relevance, we intend to build a relevance model that can capture the essence of human communication and support human-AI interaction.

5.2 Task

We design a game to simulate impromptu and sparse communication in the football example in the introduction. In this game, a player navigates the map from a starting point to a goal. Our maps are 5 by 5 as shown in Fig. 5.1. As the player wants to reach the goal as fast as possible, we assume a small cost of 1 for each step coupled with a substantial reward of 100 upon reaching the goal.

Three static monsters are randomly distributed in the 23 tiles other than the starting point and the goal. The player's objective is to avoid the monsters because stepping into each monster incurs a penalty of 100. Stepping into a monster only results in a penalty and does not end the game. Importantly, the locations of the monsters are unknown to and cannot be perceived by the player, which motivates communication. Intuitively, the player tries to reach the goal in as few steps as possible while avoiding the monsters. However, without knowing the locations of the monsters, the player can only travel through a short path.

A helper who knows the locations of the monsters tries to assist the player. To highlight the sparsity of human communication, the helper can only point to one monster in the map when the player is at the starting point. The helper can also choose not to communicate when the player is at the starting point. However, after the player starts to move, the helper cannot communicate to the player anymore.

In this immediate and perilous situation that requires on-the-spot communication, the information communicated is far less than what is needed. Three monsters have the same semantic meaning to the helper: they are all monsters. However, their locations are different, which means the pointing signal to them differ pragmatically. By assessing which monster the helper decides to point to, we want to test that a) the communication is not randomly chosen among the semantically identical signals but taking their pragmatic information other than ambiguity in RSA into consideration, and b) that the most relevant information is

communicated.

5.3 Pilot Experiment: Human Decision-making Measurement

Since we want to compare the similarity between the prediction of our relevance model and human behavior, we need to make sure that our decision-making model to be as close to human participant behavior as possible. We notice that humans do not always choose trajectories that align with the choices of a rational agent employing an MDP solver. A rational agent with a Boltzmann policy derived from MDP solver chooses actions with the same expected utility with equal probability. For example, if both going straight and turn right lead to the same destination with no difference in effort, then an MDP agent will take either of the actions with 0.5 as the probability. However, humans do not walk like this in the real world, nor do the participants operate their player figure this way. In contrast, Individuals tend to continue moving in the same direction as their initial movement (Arechavaleta, Laumond, Hicheur, & Berthoz, 2008). They seem to be a momentum when they are walking. Also, they often opt for actions that align geometrically with the goal, choosing movements that bring them closer to the destination.

This preference between actions can be modeled by an intrinsic reward. An intrinsic reward is a reward that is not physically present but reflected by the decisions made by agents. An example is exploration in navigation (Du et al., 2023). We assume that there is an intrinsic reward that drives the human participants to favor moving straight and aligning with the goal. We utilize the concept of feature expectation from inverse reinforcement learning and formulate the intrinsic reward as a linear combination of various features associated with the state (Abbeel & Ng, 2004; Bobu, Scobee, Fisac, Sastry, & Dragan, 2020). Apart from the initial reward function R, we introduce two additional features. The first feature calculates the angle between the intended action a and the previous movement direction v, $r_v = cos(a, v)$. This reward is higher when the intended action and previous movement is

pointing to similar directions. The second feature computes the angle between the intended action a and a vector w extending from the current position to the goal, $r_w = cos(a, w)$. This reward encourages the agents to move towards the goal. The illustration is shown in Fig. 5.2.

We undertook a pilot experiment to estimate the appropriate weights for these two features in influencing human decision-making during navigation. We recruited five participants to navigate the maps in our Experiment 1 but only with the walls and no monsters. There are five unique maps regarding walls. We recorded 2 trajectories for each map, resulting in 50 trajectories in total.

For each trajectory j, we calculate its reward from the environment $R(j) = \sum_t r_t$, the reward from following previous direction $v_j = \sum_t r(v, t)$, and the reward from moving towards the goal $w_j = \sum_t r(w, t)$. We fit a logit model

$$\log \frac{P(j)}{1 - P(j)} = R(j) + \alpha_1 v_j + \alpha_2 w_j \tag{5.1}$$

to learn the reward function for the agent (Train, 2009). With the trajectory collected, we estimate the parameters $\hat{\alpha}_1 = 2.497$, $\hat{\alpha}_2 = 5.077$. The reward $R + \hat{\alpha}_1 r(v, t) + \hat{\alpha}_2 r(w, t)$ is only used to predict humans' navigation policy. When evaluating the policies, we use the original reward R from the environment.

5.4 Experiment 1

In Experiment 1, we assign participants as the helper in the task, while the player was controlled by a rational AI agent with a POMDP solver. Our objective is to assess whether a relevance model can effectively capture the human decision-making process when it comes to selecting which monster to communicate within a map. We utilize our relevance model to make predictions about human choices and subsequently measure the correlation between the model's predictions and the actual choices made by humans. If the correlation shows a strong positive trend between human participants' and our relevance model's selection

probability for each monster, we can provide evidence that humans use relevance to choose what to communicate.

5.4.1 Participants

17 undergraduate and graduate students participated in this online study and were compensated with 5-dollar gift cards.

5.4.2 Maps

We generated a set of 14 distinct maps, each of which can be horizontally mirrored, resulting in a total of 28 maps. The 14 original maps are shown in Fig. 5.3. Among the 14 original maps, a player figure representing the player's position is located at position (0, 0). In the mirrored maps, the player's starting position is shifted to (4, 0).

The player can navigate the map with 4 actions: going up, down, left, or right. Diagonal movement is not allowed in the map. To introduce more diverse policies, we incorporated horizontal or vertical barriers in 9 out of the 14 maps. These barriers prevent the player from moving across them. Attempts to move across the barrier or out of the map will keep the player at its current location while paying the movement cost for one step. The game ends when the player enters the goal. The 14 maps are manually crafted with the goal of maximizing the variation of relevance in each map, guided by specific principles. First, we ensure that there is always a monster positioned off the shortest paths, like the green monster in Fig. 5.1(a). This monster's relevance is intentionally set to be very low. Second, we strategically place a monster on a cell that multiple shortest paths pass through, sometimes encompassing all the shortest paths. This monster will have a very high relevance. Not knowing its location leads to undesirable consequences of stepping into at least one monster and getting a large penalty.

Besides manipulating the relative location between the monsters and the goal, we also

alter the relationship between the locations of the monsters themselves. This manipulation involved an arrangement requiring all the shortest paths passing through the first monster to also pass through the second monster, like the monsters in Fig. 5.1(d). If a rational player is aware of the second monster, then it is very likely to avoid the first monster, too; however, in turn, if a rational player is aware of the first monster, it is still likely to step into the second monster. In this scenario, we call the first monster "blocked" and the second monster "blocking". For example, in Fig, on its own, the red monster might have high relevance. However, since all the paths through the red monster (blocked) must go through the green monster (blocking), the relevance of the red monster significantly drops compared to the green one. We assume that to thrive in this type of map, agents need the capacity of counterfactual reasoning. They need to think "what will happen if I point to that monster instead?" to point to the blocking monster instead of the blocked one.

The colors used in the examples are solely for illustration. It is important to note that the monsters are all displayed as the same color in the actual experiment, as in Fig. 5.4.

5.4.3 Design and procedure

Participants joined the experiment by accessing a link on their personal computers. The use of mobile phones and tablets was prohibited for this purpose.

Upon joining the experiment, participants were provided with a tutorial that includes an informed consent document. Then, they were given a comprehension quiz to ensure they understood the instructions properly. Prior to the formal trials, participants went through six practice trials. The practice phase included working with three maps and their respective mirror images all presented in a random order.

After completing the practice trials, participants proceeded to the main experiment which comprised of 28 trials. During each trial, the participants were presented with two maps side by side (Fig. 5.4). A large map was displayed on the left side, showing the perspective of the

helper (themselves) and revealing the locations of the monsters. In contrast, the small map on the right provided the player's view for the participant's reference. It did not show the locations of the monsters. The trials were presented to each participant in a random order, ensuring that the original and mirrored versions of the same map were not shown in two consecutive trials. During the trials, participants could assist their partner by highlighting a monster on the large map. They did this by double-clicking on the selected monster. The participants' choice was then recorded.

After the participants confirmed their choice of highlighted monster, a nine-point Likert scale appeared on the screen. This scale prompts participants to rate the perceived helpfulness of their signal, ranging from 1 (not helpful) to 9 (very helpful). The rating was also recorded for analysis. Upon receiving the highlighted location from the participant, the AI player updates its belief, treating the marked location as a monster. It then updates its policy based on this new information and proceeds to navigate the map accordingly. The navigation route was visually presented to the participants as an animation. Once the player successfully reached the goal, a feedback box appeared on the screen. This feedback included details about the partner's performance, such as the number of steps the player took, the number of monsters they encountered, and the score they received. After reviewing the feedback, participants proceeded to the next trial. The participants took an exit survey after completing all the trials.

5.4.4 Results

Choice of pointing. For each map across participants, we compute the probability of each monster being selected. This probability can be denoted as p_{human}. Initially, we put all the monsters from all maps together, treating all monsters independently to find the correlation. We predict the probability of choosing the monster $p_{relevance}$ with our relevance model. The participants' choice of pointing can be linearly predicted by a softmax of relevance, as shown in Fig. 5.5. The correlation between p_{human} and $p_{relevance}$ is $r(40) = 0.695, p < .005$. When

using $p_{relevance}$ to linearly predict p_{human} with the equation $p_{human} = \beta_0 + \beta_1 p_{relevance}$, the regression coefficient is $\hat{\beta}_1 = 1.245, t(40) = 6.111, p < .001, R^2 = 0.483$.

We then specifically look at the most relevant target within each map. The higher the relevance, the more likely it is to be picked by the participants. The probability of choosing the most relevant monster can be predicted by the exponential of its relevance (Fig. 5.5). Among all the maps, even the monster with the highest relevance observes a large variation in relevance. In three maps, the monster with the highest relevance still does not have a very high relevance (less than 10). In those same maps, we expect the probability of the monster with the highest relevance being chosen to be lower than in other maps, since the most relevant monster is only slightly better than the other two monsters.

We calculate the logit of p_{human} for the most relevant monster on each map, $logit(p_{human})$ $= \log \frac{p_{human}+\epsilon}{1-p_{human}+\epsilon}$, with a hyperparameter arbitrarily chosen as $\epsilon = 0.001$. The correlation between relevance and $logit(p_{human})$ is $r(12) = 0.899, p < .001$. If we use relevance to linearly predict the logit of p_{human} with the equation $logit(p_{human}) = \beta_0 + \beta_1 relevance$, the regression coefficient is $\hat{\beta}_1 = 0.053, t(12) = 7.113, p < .001, R^2 = 0.808$.

To further analyze the act of pointing to monsters, we study the choice of monsters that fall into the four categories introduced in the stimuli section. Participants chose the monsters off all shortest paths in only 3.78% of the trials. It is very close to 0 as predicted by the relevance model. In the 7 out of 14 maps where the relevance of the most relevant monster is high (> 40), participants chose the most relevant monster in 83% of the trials. For the maps where blocking occurs, participants chose the most relevant monster in 82.3% of the trials. In 13.7% of the trials, they chose the second most relevant monster, the blocked one.

Helpfulness self-rating. We also examine the participants' self-rating of their pointing's helpfulness. When analyzing helpfulness self-rating data, we excluded data from one participant who reported not to correctly understand the rating. When the participants pointed to the most relevant monster, they rated the pointing as more useful. We use whether the participant chose the most relevant monster ($M = 7.11, SD = 1.89$) or not ($M =$

$6.16, SD = 2.10$) to predict their self-rating of helpfulness and fit a fixed-effect linear model $rating = \beta_0 + \beta_1 I(highest\ relevance) + \alpha_i participant_i$. $\hat{\beta}_1 = 0.786, p < .001, R^2 = 0.319$.

Self-report strategies. We check the participants' strategies in the game, which were self-reported after the experiment. 5 out of 11 participants who reported their strategy mentioned that they took the player's perspective of making decisions. 9 out of the 11 participants mentioned action prediction and evaluation with phrases such as "shortest path". One participant reported counterfactual reasoning. All the participants' answers of their strategies are shown in Table 5.1.

5.4.5 Discussion

The linear relationship between relevance and the logit of the participants' choice probability indicates that humans utilize utility-based relevance when deciding which information to communicate. Evidenced by the ratings, humans possess an awareness of the degree of helpfulness associated with each piece of information and thus tend to offer the most beneficial information. The participants' self-reported strategies clearly show that humans engage in action prediction and evaluation when assessing relevance, and they also incorporate counterfactual reasoning.

5.5 Experiment 2

We turn our task in Experiment 1 to a language game and use the state-of-the-art large language model (LLM) GPT-4 (OpenAI, 2023) as a participant. GPT-4 was provided with a description of the game and was asked to choose one of the three monsters to point to.

5.5.1 Stimuli

An example of the prompt we provided is: "You are playing a two-person game. A player navigates a 5×5 grid map to a goal. In the map, there are three monsters invisible to the player. The player can see the location of the goal. When the player reaches the goal, the game ends. The player can also see the walls on the map that they cannot go through. The walls are between two cells in the map. The player going one step costs 1 dollars. Reaching the goal grants 100 dollars and running into each monster costs 100 dollars. We want to get the highest reward. You play as a helper who can inform the player of the location of only one monster. Do you understand the game?"

After GPT-4 repeated the correct rules, we entered the second prompt: "If the player starts from (0, 0), the goal is (2, 4), no walls are in the map, the monsters are at (1, 2), (2, 0), and (4, 2). Which monster will you tell the player?"

We recorded GPT-4's responses and compared them with the human data collected from Experiment 1.

5.5.2 Results

Choice of pointing. We compared GPT-4's choices with the human participants' from Experiment 1.

40.8% of human participants' choices are consistent with the choices from GPT-4. In comparison, 60.9% of human participants' choices are consistent with the choices from our relevance model. Humans' choices are more consistent with our relevance model than GPT-4's prediction, $z = -6.223, p < .001$.

We analyzed the pointing to each monster in the four categories in more detail (Fig. 5.6). In 13 of the 14 maps, GPT-4 will not point to the monster off any shortest path, with the exception shown in Fig. 5.1(b). GPT-4 chose the orange monster, which is closest to the starting point, but not on any shortest path. No human participants chose this monster.

This may result from an inaccuracy in GPT-4's planning.

In the seven maps where one monster is very relevant, GPT-4 only chose the most relevant monster in 29.6% of the trials, far from the proportion in human participants (83%).

In the maps where one monster blocks another, GPT-4 chose the most relevant monster with a probability of 25%. In the other 75% of the trials, it chose the blocked monster. For example, in Fig. 5.1(d), where the the red monster is blocked by the green one. GPT-4 points to the red monster, which is rarely consistent with human choice. As we speculated when we designed the maps, this outcome potentially indicates that GPT-4 lacks the ability for counterfactual reasoning to some extent.

Helpfulness rating. We also repeated the study of the helpfulness self-rating in Experiment 1. Since we cannot obtain self-rating data from GPT-4, we used the helpfulness rating data from Experiment 1 as a source for estimation. We split the data in Experiment 1 into two conditions: if the participants had the same choices as GPT-4 or if the participants chose differently than GPT-4. We fit a fixed-effect linear model to find no difference in helpfulness rating if participants chose the same ($M = 6.60, SD = 2.02$) or different ($M = 6.83, SD = 2.03$) monster than GPT-4. $rating = \beta_0 + \beta_1 I(GPT4) + \alpha_i participant_i$, $\hat{\beta}_1 = -0.155, p = 0.34, R^2 = 0.253$.

Self-report strategies. GPT-4 also provided intact and reasonable strategies that were similar to the human participant reports in Experiment 1. The answers clearly involve action prediction and evaluation. GPT-4 assumes one trajectory for the player and points to a monster on the trajectory. However, it does not consider all of the player's possible policies. We see human participants report this strategy too, but their choices of pointing to monsters are different from GPT-4.

5.5.3 Discussion

The consistency of GPT-4's choices with human participants is lower than the consistency of relevance model with participants, but the strategy report from GPT-4 is very similar to humans. The strategy reports suggest that GPT-4 encompasses the concepts of action prediction and action evaluation, showing a certain level of capacity for theory of mind and planning. However, the action-planning in GPT-4 is not always reliable. GPT-4 may just mimic human thinking processes on a surface level.

5.6 Experiment 3

We flip the role of the human participant and AI in Experiment 1. We assigned participants as the player in our task, while the helper is a rational AI who communicates either with our relevance model or a heuristics model. The heuristic model points to the monster closest to the starting point as measured by Manhattan distance. For example, in Fig. 5.1(c), the heuristic model will point to (0, 2). In 10 out of the 28 maps, the heuristic model points to the same monster as the relevance model. It is worth noting that this experiment was conducted before the launch of GPT-4. In the future, we are interested in involving GPT-4 as a model that provides pointing to compare performances.

Our goal is to test a) whether a relevance model can contribute more to task performance than a heuristic model and b) whether the help from a relevance model is better received by human participants.

5.6.1 Participants

20 undergraduate students participated in this study and were compensated with course credits.

5.6.2 Stimuli

The maps were the same as in Experiment 1.

5.6.3 Procedure

Participants entered the experiment room and opened the link to the experiment on a computer. Then, participants went through a practice phase similar to Experiment 1. After the practice phase, the participants started the formal 28 trials, arranged in the same manner as Experiment 1. In each formal trial, participants saw the map without monsters. After 3-5 ms, a highlight on a cell showed the monster marked by the helper. The participant then controlled the keyboard to navigate the map to the goal. The reward gained by the participants was recorded.

After the participant reached the goal, all the monsters were revealed on the map. A feedback box appeared, showing the participant how many steps they took, how many monsters they ran into, and the score they received. The feedback box also contained a nine-point Likert scale asking the participants to rate the helpfulness of the signal (from 1-not helpful to 9-very helpful). The rating was also recorded for analysis.

5.6.4 Results

One participant was excluded from the analysis, because all their helpfulness ratings were marked as 1.

Reward. We compare the reward from relevance model and the heuristic model. A paired t-test shows that the participants achieve higher reward when receiving help from the relevance model ($M = 53.744, SD = 17.980$) than when receiving help from the heuristic model ($M = 43.255, SD = 12.571$). $t(18) = 2.448, p < .05$.

Specifically, we compare the reward received in the 18 maps where the relevant helper

and the heuristic helper point to different monsters. In these 18 maps, a paired t-test shows that the participants achieve higher reward when receiving help from the relevance model ($M = 58.503, SD = 19.631$) than when receiving help from the heuristic model ($M = 40.327, SD = 16.430$). $t(18) = 3.599, p < .005$.

Helpfulness rating. We compare the ratings received from the participants to the relevance model and the heuristic model. A paired t-test shows that the participants give higher ratings to the help from relevance model ($M = 5.530, SD = 1.547$) than to the help from the heuristic model ($M = 5.056, SD = 1.47$). $t(18) = 3.549, p < .005$.

Furthermore, we compare the rating received by participants in the 18 maps where the relevant helper and the heuristic helper point to different monsters. In these 18 maps, a paired t-test shows that the participants give higher ratings to the help from the relevance model ($M = 5.650, SD = 1.724$) than to the help from the heuristic model ($M = 4.930, SD = 1.421$). $t(18) = 2.948, p < .01$.

5.7 Discussion

Through our experiments, we present concrete evidence that impromptu human communication within cooperative contexts aligns with the principle of utility-based relevance. In our first experiment, we observe that the choices made by human participants regarding which monster to point to can be accurately predicted by our relevance model. Notably, the consistency between the participants' choices and the model's predictions remains high across various types of monsters, including those with very low relevance, very high relevance, blocked paths, and those blocking paths.

These findings suggest that when individuals decide what information to convey, they understand the relevance associated with each piece of information. Then they rationally choose the relevant ones. RSA proposes that each utterance in communication carries a utility based on how it alters the listener's belief towards the true belief in a reference game

(Frank & Goodman, 2012). In our experiment, we expand the concept of utility to relevance: the utility of a communicative signal within a task can be defined by how it enhances the listener's performance. Our results are in line with recent research on human communication (T. R. Sumers, Hawkins, Ho, & Griffiths, 2021).

In Experiment 3, we observe that communication generated by AI integrated with a relevance model yields enhanced performance and receives higher ratings in terms of perceived helpfulness. We aspire that our model can serve as a source of inspiration for designing AI systems that offer both impromptu and efficient interactions with humans.

An AI system equipped with a relevance model only requires planning capacity grounded in decision-making theory to effectively provide concise assistance to humans. This approach could potentially alleviate the intensive training needed for AI to communicate in human natural language or develop a new language (Lazaridou & Baroni, 2020; Havrylov & Titov, 2017). Additionally, it may circumvent the necessity of AI sharing its entire observation (Reddy et al., 2021) with humans. Instead it can pick the most relevant information in the observation to avoid excess load of information in the cases of large observations, such as with images or videos.

Comparing the strategy reports from human participants in Experiment 1 and GPT-4 in Experiment 2 reveals a noteworthy trend: most individuals engage in both action prediction and action evaluation when calculating relevance, a pattern that GPT-4 is capable of capturing. This observation suggests that human communication heavily relies on theory of mind and planning (Jara-Ettinger et al., 2016), capacities which GPT-4 seems to possess (Sap et al., 2022). However, there is a distinction: GPT-4 is more inclined to simulate thought rather than genuinely think (Mahowald et al., 2023).

Our results highlight a substantial difference between the choices made by GPT-4 and those by human participants when it comes to selecting information. This prompts our hypothesis that LLMs primarily learn from frequently expressed language, which only represents a fraction of human thought processes. This is evident in scenarios such as our

maps with monster blocking, which require counterfactual reasoning—an ability intrinsic to humans from a young age (Buchsbaum, Bridgers, Skolnick Weisberg, & Gopnik, 2012). While most human participants correctly chose the blocking monster, only one explicitly mentioned counterfactual reasoning in their strategy report. In contrast, GPT-4 often chose the blocked monster, showcasing its deficiency in counterfactual reasoning. This could be tied to the infrequent mention of counterfactual reasoning in human language, making it a challenge for LLMs to simulate such capacities.

In the thriving trend of reinforcement learning with human feedback (Ouyang et al., 2022), it is pivotal to appreciate the value of incorporating behavioral data into model training and conducting data analysis guided by theories in cognitive science.

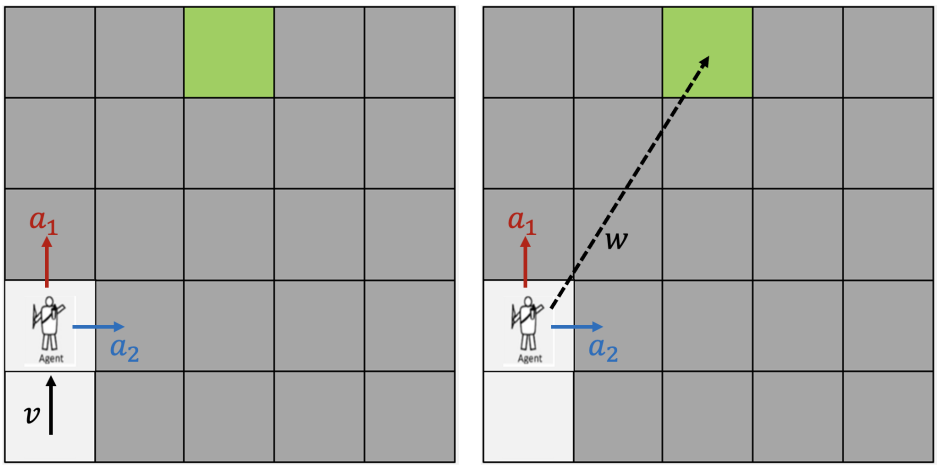

Figure 5.2: **Features used to model participants' innate reward when selecting trajectory. Left:** participants tend to choose a_1 which aligns with momentum v; **Right**: participants tend to choose a_1 which directs closer to the goal.

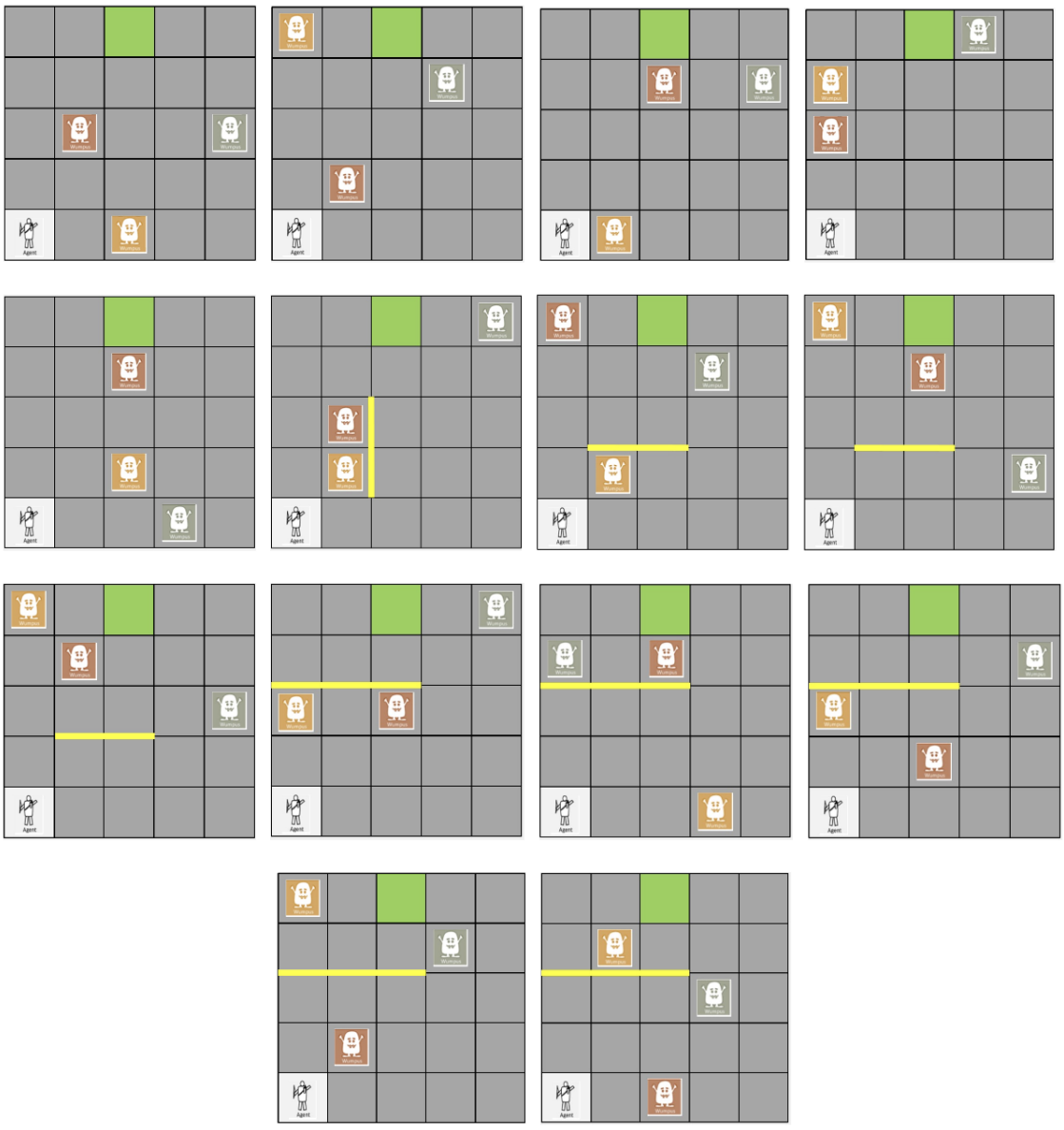

Figure 5.3: **List of all maps**

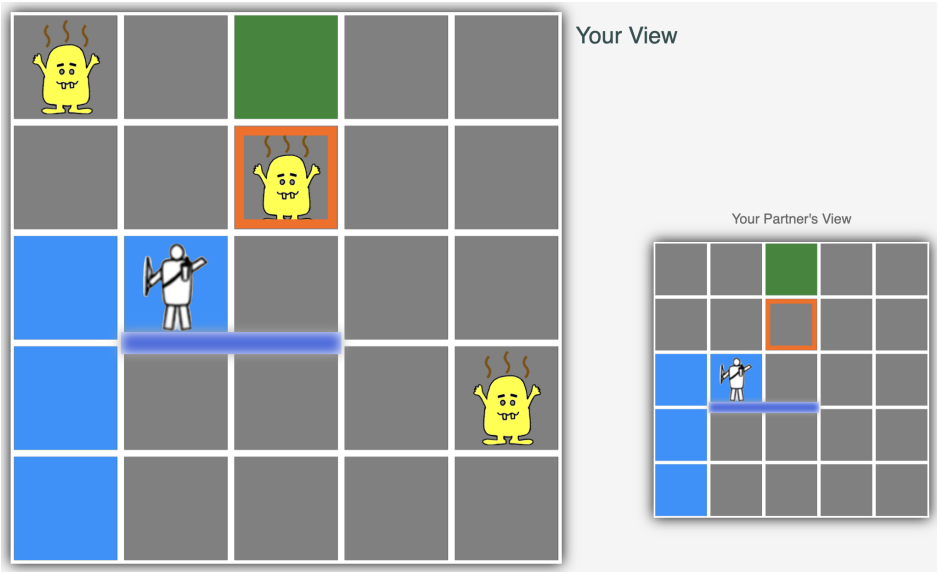

Figure 5.4: **A frame in display of Experiment 1:** after participant selects monster at position (2, 3) and completes helpfulness self-rating, the player figure moves to the goal.

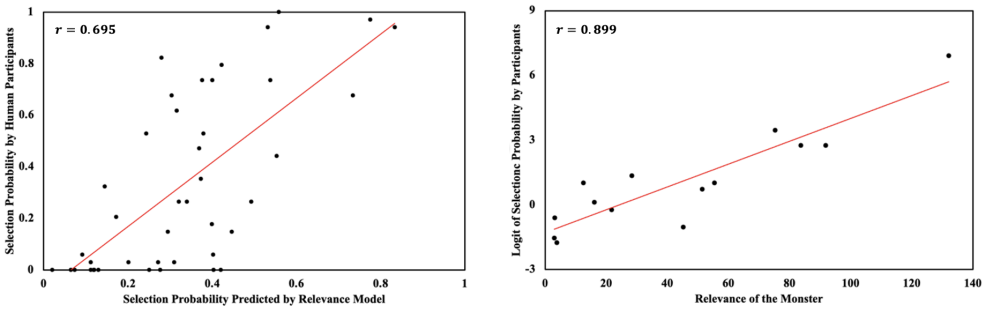

Figure 5.5: **Linear relationships of human participant selection probability and relevance. Left:** Predicted probability from relevance vs. human selection probability; **Right:** Relevance vs. logit of human selection probability

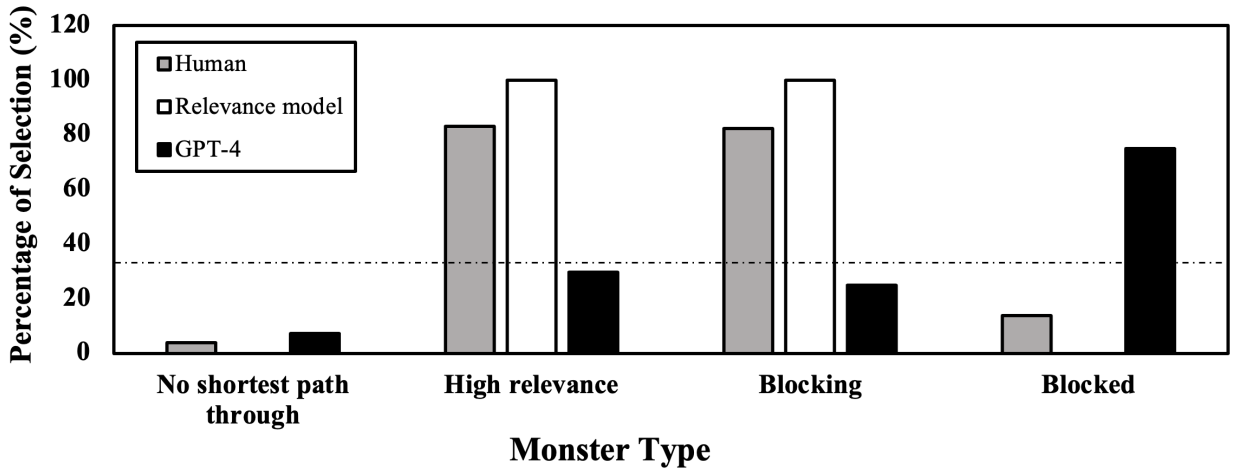

Figure 5.6: **Selection Probability of human, relevance model and GPT-4 for types of monster.** Dashed line: chance probability (33.3%)

Try to interpret my partner's strategy, and find out the monster which would be most helpful to point out in respond to the partner's strategy.
1. find all possible shortest path from partner's point of view; 2. signal the monster that blocks the most number of paths
If the monster is right beside the starting point or the goal, I point to that monster. If not, I plan a shortest path myself and see which monster impacts the path the most.
If two monsters are next to each other, I'll choose one of them. If a monster is on the shortest path that the partner is likely to pass, I will point to it and make the partner circumvent to a longer and safer path.
Try to think in my partner's shoes and determine which monster is the most dangerous if I were the partner.
Plan a path in my mind as a player and if that path passes through a monster, I click on that monster.
Choose three shortest paths that I prefer and see which monster impacts them the most.
I tried to use the pointing signal to warn the partner of a direction that could potentially harm its utility the most. It's a pity that I could not point to an empty space since sometimes pointing to an empty space does a better job on that.
If I feel two points are both important, I will point to one but give myself a low rating. Because I feel pointing to one is not enough, it may also hit the other one.
I choose one point. If I say my choice is helpful, that means people can avoid the shortest paths through it. If it's not helpful, then there is another point on the shortest path.
I imagine I'm the partner. I will choose one route to the goal that I like best. But this route almost always hits monsters. Back to the perspective of myself, I know the first monster on this route and I will point it out.

Table 5.1: Strategies reported by participants in Experiment 1.

CHAPTER 6

Modeling the Timing for Human Communication

6.1 Background

Imagine you are going to a neighborhood. You want to find a house among four possible houses in the neighborhood. You know the locations of the four houses, but do not know which one is your target. You can ask for directions from a friendly local resident. When do you want to get help? We certainly do not want to ask before we arrived at the neighborhood because what we get will be most likely a general direction: go that way to the neighborhood. We do not want to ask too late, for example, after we have already passed a possible target house because we probably need to go back. We do not ask for help through communication all the time, but only when we expect that the response helps.

Similarly, imagine in the above scenario, you are an assistant monitoring the protagonist navigating to the neighborhood. However, you can only send one message that is from the four directions: North, East, South, and West. When do you want to send the message? The same concerns stand out as in the above question asking scenario. If we send out the direction too early, it will not be clear because it is very likely that more than one possible target is in that direction. If we send out the direction too late, the protagonist may have already entered a wrong place, causing some troubles.

From the two examples, it is obvious that we do not communicate all the time. Instead, we only communicate when it is a good timing: when the communicative signal can be clear, straightforward, and relevant. In this chapter, we aim to study the temporal characteris-

tics of relevance-based communication. We derive when it is an optimal time window to communicate.

6.1.1 Sparse Human Communication

Human communication is sparse. In engineering, the sparsity of the communication is attributed to the sparsity of stimuli or reward in the environment, for example in aviation, planes do not observe many stimuli through its sensor, or the sparsity of the communication channel, like communication between faraway buildings or cities (Tong, Zhang, Wang, Huang, & Debbah, 2021). However, in human communication, the sparsity may result from the fact that people do not need to say that much. In communication, we have a transparent common knowledge, information in which we do not need to communicate about. Also, under the Gricean communicative axioms (Grice, 1975), the communicators are constrained by the commitment to be clear, straightforward, and relevant. If we cannot send out a good signal, our partner in communication may blame us for not being capable to communicate efficiently.

In multi-agent reinforcement learning, the objective of sparse communication is to minimize the total number of bits communicated while maximizing team task performance (Karten, Tucker, Kailas, & Sycara, 2023). It can be formulated as a minimax problem for all the agents i:

$$\max_{\pi:S\to A\times M} E\left(\sum_t \sum_i \gamma R(s_t, a_t)\right), \quad \text{subject to } \min E_{m\sim\pi}(s(M)) \tag{6.1}$$

Here T and R represent the transition and reward functions respectively, $a \in A$ is the action of the agent, $m \in M$ is the message sent out by the agent, π is a joint policy of physical actions a and communicative actions m, and $s(M)$ measures the message flow in the information transmission process. This minimum problem can be rephrased by introducing

a Lagrangian,

$$\max_{\pi:S\to A\times M} E\left(\sum_t \sum_i \gamma R(s_t, a_t) - \lambda s(m_t)\right), \lambda \geq 0. \tag{6.2}$$

We adopt this minimax perspective but keep the setting of a player and a helper in a navigation scenario. We take an approach of deriving the optimal timestep to communicate by a) restricting the number of communications to 1, and b) converting the optimization problem as a Markov decision process (MDP) or a partially observable Markov decision process (POMDP). We will start from introducing the fundamentals of formulation, followed by derivation and simulation experiments.

6.2 Formulation

6.2.1 MDP and POMDP

A discrete time partially observable Markov decision process (POMDP) can be represented by a tuple S, A, Ω, T, R, O. S represents all the possible states, A represents all the possible actions, Ω represents all the possible observations. T represents the transition process between the states, O represents the sensory process of how agents receive observations from the states, R represents the reward function which describes how agents receive reward from the states. A Markov decision process (MDP) is a special case of POMDP whose S and Ω are the same and O is an identity map (Montufar, Ghazi-Zahedi, & Ay, 2015). Therefore, we can represent an MDP with a tuple of four components: (S, A, T, R). A more detailed introduction of POMDP and MDP can be found in Chapter 2.

A policy π describes how the agent selects actions. It is a probability distribution over actions based on the current state $p(a_t|s_t)$ for MDPs, and a probability distribution over actions based on the current belief state $p(a_t|b_t)$ for POMDPs. The goal for solving MDPs and POMDPs is to find a policy that maximizes the cumulative expected reward.

6.2.2 Formulation of the cooperative direction guide

In our formulation of the house locating problem as introduced at the start of the chapter, we assume two agents: one receiver and one sender. The receiver can only navigate physical the world with no communication, while the sender can only help the receiver by sending out at most 1 communicative signal without the capacity to perform any physical actions. The different target houses represent different reward structures. The sender knows the reward structure of the world and the location of the receiver. We assume that the receiver has a belief b_P of where the reward is, over all possible states of the world. In this case, the receiver navigates the world with a POMDP framework. The belief reduces to a goal g if the receiver is certain of where the goal is: The belief is 1 in the state g and 0 for all other states. We start from the assumption that the sender knows this belief or goal b_P and show that the communication process is an MDP for the sender. Then we remove the assumption, let the sender and show that communication is a POMDP for the sender.

6.3 Deciding When to Point Only Once Is An MDP When Belief Is Known

For a receiver navigating the world, we can see him as an agent with single-agent POMDP. The components for this POMDP can be shown in a tuple S, A, Ω, T, R, O. In this POMDP, the state space is all the possible states in the environment, the action space is all the possible actions that he can do to navigate the world, the observation space is all the observations that he can get. Take the guided Wumpus hunt (Chapter 4) as an example, the state space is a tuple (L_P, L_W) representing the location of the player and the Wumpus; The action space is {move left, move right, move up, move down, shoot left, shoot right, shoot up, shoot down}; The observation space is {stench, nothing}.

In this process, the transition function, reward function, and observation function repre-

sent the mechanism of interactions between the agent and the environment. The transition function shows how the state changes according to the actions of the agent: when the player chooses to move, L_P will change to one further tile on the intended direction while L_W does not change; When the hunter chooses to shoot, the game ends. The reward function shows how the agent gains reward from the environment: when the hunter moves without running into the Wumpus, he will pay a movement cost; When the hunter runs into the Wumpus or misses the shot, he will pay the penalty; When the hunter hits the Wumpus, he will gain the 100 points. The observation function shows how agent receive the observation from each state of the environment: When the hunter is in a tile adjacent to the Wumpus, he has a large probability to smell the stench while a small probability to smell nothing; When the hunter is in a tile far away from the Wumpus, he has a small probability to smell nothing and a large probability to smell the stench. The components in S, A, Ω, T, R, O altogether helps the receiver derive the optimal policy π.

In the communication process, the action of communication is also a decision-making process. We argue that this can be modeled as an MDP if the sender knows the receiver's belief, policy, and belief transition after receiving each communicative signal.

Here we model the communication process as two sequential decision-making processes alternate in the timeline. In the beginning of each timestep, the sender has a belief $b_{H,t}$ and the receiver has a belief $b_{P,t}$. In each step, the sender first decides how to communicate and takes the action u; the receiver's belief changes to b_P^u in response to the communication. Then the receiver decides how to act in the environment and takes the action a. He receives the reward r and observation o. His belief transitions to b', which is the initial belief $b_{P,t+1}$ for the next step. In turn, the receiver's action, observation, and state transition may be observed by the sender as observations, from which she can infer the transition in the receiver's beliefs.

Proposition 1 Assume that the sender knows the state, the receiver's belief, policy, and belief transition after receiving each communicative signal. If the sender can only decide whether to communicate at the current timestep and can never communicate in the future

timesteps, its one-step communication decision can be represented as an MDP (on the belief space) for the sender.

Proof of Proposition 1. The sender H knows the state s, and a set of signals C ($silence \in C$) available. If the receiver P has a POMDP represented by (S, A, Ω, T, R, O) and a policy $\pi(a|b)$ for each possible belief b, then the sender can evaluate the utility of every possible belief $b \in \Delta_S$ owned by the receiver. Δ_S represents the set of probability distributions over S. The utility can be represented by a function $U : \Delta_S \to \mathbf{R} : p(s) \mapsto U_{b_H}(b)$. We also add a belief state end to the belief state space to represent that the sender can no longer decide to communicate. If the signaler also has access to how the receiver changes his mind according to signals, represented as $T_b : \Delta_S \times C \to \Delta_{\Delta_S}; b, u, b' \mapsto I(b' = end)$.

From the sender's perspective, the reward for the belief transition can be calculated as $R_b : \Delta_S \times C \times \Delta_S \to \mathbf{R}; b, u, b' \mapsto [U_{b_H}(b^u) - c(u)]I(b \neq end)$ since the receiver follows a single-agent POMDP with a belief b^u after the receiving the communicative signal u. $c(u)$ represents the cost of sending the signal u. The one-step communication of the sender can be represented as an MDP (Δ_S, C, T_b, R_b).

We can also represent $V_{H,0}(b) = U_{b_H}(b^u) - c(u)$, where 0 represents after this decision, no communication can be attempted. $V_{H,0}(b)$ is the maximum expected cumulative reward of the receiver evaluated from the sender's perspective after receiving the communicative signal because it is a single-agent POMDP.

Proposition 2 Assume that a) the sender knows the state, the receiver's belief, policy, and belief transition after receiving each communicative signal, and b) the sender can only communicate once and only silence in other timesteps, then the communication process can be represented as an MDP for the sender.

Proof of Proposition 2. The sender H knows the state s, and a set of signals $C = \{utterances, silence\}$ available, where C is a set of communicative signals. The receiver P has a POMDP represented by (S, A, Ω, T, R, O) and a policy $\pi(a|b_P)$. At each timestep,

the sender has two choices: to communicate or to keep silent. If the sender chooses to communicate, then she cannot communicate anymore. As stated in the proof of Proposition 1, she cannot communicate anymore, so the reward received is $V_{H,0}(b_P) = U_{b_H}(b_P^u) - c(u)$. If she chooses to keep silent, the receiver's belief will change to b' with a probability $P(b'|b,u)$ when the sender can choose to point in the next timestep. Here $P(b'|b,u)$ models the change of receiver's belief caused by two observations: one is the sender's communication u, the other is observation $o \in O$ received from the environment.

$$P(b'|b,u) = \sum_o P(b'|b,u,o)P(o|b,u) \tag{6.3}$$

$$= \sum_o P(b'|b^u,o)P(o|b^u) \tag{6.4}$$

$$= \sum_o \sum_a I(b' = SE(b^u,a,o))P(o|b^u,a) \tag{6.5}$$

$$= \sum_o \sum_a I(b' = SE(b^u,a,o)) \sum_s P(o|s',a)P(s',a,s) \tag{6.6}$$

The state transition and observation o are generated from the physical state of the world instead of the beliefs, but it can serve as stochasticity in the transition function.

Meanwhile the receiver (the same for the sender) will receive the one-step reward $r(s) = E_{a\sim\pi(a|b_P)} \sum_s' R(s,a,s')T(s,a,s')$ evaluated by of the sender.

Therefore, the communication of the sender can be represented as an belief-space MDP (Δ_S, C, T_1, R_1). The transition function is $T_1 : \Delta_S \times U \times \Delta_S \to \Delta_{S,U,S} : b,u,b' \mapsto [I(b' = end)I(u \neq silence) + P(b'|b,u)I(u = silence)]I(b \neq end) + I(b = end \text{ and } b' = end)$. The reward function is $R_1 : \Delta_S \times U \times \Delta_S \to \mathbf{R} : b,u,b' \mapsto [V_{H,0}(b)I(u \neq silence) + r(s)]I(b \neq end)$.

With the formulation of MDP, we can solve for a value function $V_{H,1}(b)$ for each belief possible when the sender can only send out one communicative signal. We can replace $V_{H,0}(b)$ with $V_{H,1}(b)$ if we want to derive the reward function and $V_{H,2}(b)$ for the MDP with two possible communications, and so on.

6.4 Application: Direction Guide

To test our sparse communication model, we augmented the maps from (C. Baker et al., 2011) to design a task called Direction Guide. Since belief state MDPs are difficult to solve, we simply the problem by making a few assumptions: a) the receiver does not receive any observations other than the location of himself, and b) the receiver's belief does not change if no communication occurred.

In Direction guide, a player navigates a 7×7 grid world to collect a reward. There are four possible locations in the world, but only one contains the reward. The player can only move up, down, left, or right in the grid world with a small moving cost for each step. As a result of the action, he will move deterministically in the direction of the moving actions, stay put if hit a wall or a boundary, or end the game when the player enters one of the four possible reward locations. He may or may not have a belief of where the reward is. The player obtains the reward if the reward is at this location while get a penalty if the reward is not at this location. When navigating the world, the player will receive no observations other than his current position.

In the Direction guide, the player can formulate the problem as a POMDP, where the state space is a tuple of the location of the player and the location of the reward $S = (L_P, L_R)$. The actions are $A = \{move\ up, move\ down, move\ left, move\ right\}$. Since there are no observations other than the location of the player himself, we can reduce the problem to an rMDP (Littman et al., 1995).

A guide knowing the location of the reward tries to help the player. She can only help by communicating to the player once in the entire game, while her communication signals are limited to the four directions: up, down, left, right.

6.4.1 Formulating the Direction Guide with rMDP

In an rMDP, the player knows his location L_P, so the uncertainty in the environment is the location of the reward L_R. Four possible locations of the reward introduce four possible reward structures $R_{l_R}(l_P, a, l'_P)$ when $L_R = l_R$. His belief can be represented as $b_P = P_{L_P,L_R}(l_P, l_R) = P_{L_R}(l_R)I(L_P = l_P)$. Since no observations about the reward structures can be obtained in the navigation, we can see the navigation problem as a weighted sum of four different MDPs, each with its own reward function $R_{l_R}(l_P, a, l'_P)$. These MDPs have the same state spaces L_P, the same action spaces A, and the same transition functions $T(l_P, a, l'_P)$.

With a formulation of rMDP, we can solve the rMDP by averaging the reward function with belief of the reward as weights,

$$R_{b_P}(l_P, a, l'_P) = \sum_{l_R} R_{l_R}(l_P, a, l'_P)P(l_R|b_H). \tag{6.7}$$

Then we can use an MDP solver, for example, the value iteration algorithm to obtain the optimal policy. R_{b_P} is the reward structure used by the player to navigate the world. With the averaged reward function, we can write down the rMDP of the player as a tuple The sender knows the true reward structure $R_{b_H} \in \{R_{l_R}|l_R \in L_R\}$.

6.4.2 Relevance Iteration

After we solve the rMDP, the sender H can predict the policies of the player $\pi(a|b_P)$, as she has access to the player's belief b_P, hence evaluate the player's belief. To evaluate the utility of each communicative action, the sender needs to simulate the belief change after receiving each signal. We start by assuming that the belief change only depends on the location and belief of the reward structure in the current timestep, then loosen this assumption to that the belief changing depends on the trajectory so far and current belief of the reward structure.

Assuming that the belief change only depends on the location and belief of the reward

structure in the current timestep, the sender can simulate the distribution $P(b'_P|b_P, u)$ over all possible beliefs of the player b'_P after she sends out a signal u. The sender can then evaluate each of the possible beliefs with $U_{b_H}(b'_P)$. The receiver's expected utility is the relevance of the signal,

$$Rel(u|b_P) = \sum_{b'_P} U_{b_H}(b'_P) P(b'_P|b_P, u). \tag{6.8}$$

This is also the reward received by the sender if she chooses to send out communication signal u, because she cannot send out any more communicative signals.

If the sender chooses to be silent, she will observe the player play this step until she gets another chance to communicate in the next timestep. The state transition that she expects to observe is: $\sum_a \pi(a|b_P) T(l_P, a, l'_P)$. The reward received until the next chance to communicate is $\sum_a R_{b_H}(l_P, a, l'_P)\pi(a|b_P)T(l_P, a, l'_P)$.

The communication decision-making problem for the sender can then be transformed to an MDP (S_H, C, T_H, R_H). The state space $S_H = L_P \cup \{end\}$ is all the possible locations for the player and the end state because the transition and reward functions for the sender only depend on the current locations of the player. The action space C is all the communication signals available along with the option to keep silent. The transition function is

$$T_H(s, u, s') = \begin{cases} \sum_a \pi(a|b_P) T(l_P, a, l'_P) & u = silence, s \in L_P, s' \in L_P \\ 1 & u \neq silence, s \neq end, s' = end \\ 0 & o.w. \end{cases} \tag{6.9}$$

The reward function is

$$R_H(s, u, s') = \begin{cases} \sum_a R_{b_H}(l_P, a, l'_P)\pi(a|b_P)T(l_P, a, l'_P) & u = silence, s \in L_P, s' \in L_P \\ Rel(u|b_P) & u \neq silence, s \neq end, s' = end \\ 0 & o.w. \end{cases} \tag{6.10}$$

After defining the components in the MDP, we can use MDP solving techniques to derive the optimal actions, which is whether the sender wants to communicate or not at this timestep.

6.4.3 Relevance iteration on trajectory rMDP

In communication, it is common when a decision or inference is made after observing a series of observations or signals. An example is when the player does not know the existence of the helper at first. When the helper sends out the communicative signal, the player realizes that there is a helper. However, he will not assume that the helper appears at this timestep, but very likely that the helper chooses not to communicate in the past timesteps. The player will also infer from the past timesteps why the helper did not communicate. In this scenario, the change of belief does not only depend on the current state of the receiver, but the whole trajectory instead.

If the belief change only depends on the belief of the reward structure and the trajectory so far, the Markov property is violated as the reward function is no longer defined on a state or a component of the state. In this case, we need to redefine the player's MDP to keep the Markov property.

We can define the rMDP (J, A, T_j, R_j) with all trajectories from timestep 1 to infinity as states $J = \{j : L_P^{1-\infty}\}$. The beliefs of the player are in the belief space $b_P = P_{J,L_R}(j, l_R) = P_{L_R}(l_R)I(J = j)$. The actions are $A = \{move\ up, move\ down, move\ left, move\ right\}$. The transition function T_j and the reward function R_j can be adapted from the rMDP transition function T and reward function R.

$$T_j(j, a, j') = T(l_P, a, l'_P)I(l_P \textit{ is the last state in } j \textit{ and } j' = (j, l'_P)), \tag{6.11}$$

$$R_j(j, a, j') = R_{b_P}(l_P, a, l'_P)I(l_P \textit{ is the last state in } j \textit{ and } j' = (j, l'_P)). \tag{6.12}$$

Evaluated by the sender, the relevance of a communicative action u is:

$$Rel(u|b_P) = \sum_{b'_P} EU_{b_H}(b'_P)P(b'_P|b_P, u). \tag{6.13}$$

This is also the reward received by the sender if she chooses to send out communication signal u.

If the sender chooses to be silent, she will observe the player play this step until she gets another chance to communicate in the next timestep. The state transition that she expects to observe is: $\sum_a \pi(a|b_P)T_j(j, a, j')$. The reward received until the next chance to communicate is $\sum_a R_{j,b_H}(j, a, j')\pi(a|b_P)T_j(j, a, j')$, where $R_{j,b_H}(j, a, j')$ is denoting the reward function evaluated from the sender's perspective $R_{b_H}(l_P, a, l'_P)I(l_P \textit{ is the last state in } j \textit{ and } j' = (j, l'_P))$.

Then we can represent the sender's MDP (S_H, C, T_H, R_H) with trajectories as states. $S_H = J \cup \{end\}$.

$$T_H(s, u, s') = \begin{cases} \sum_a \pi(a|b_P)T(j, a, j') & u = silence, s \in J, s' \in J \\ 1 & u \neq silence, s \neq end, s' = end \\ 0 & o.w. \end{cases} \tag{6.14}$$

$$R_H(s, u, s') = \begin{cases} \sum_a R_{b_H}(j, a, j')\pi(a|b_P)T(j, a, j') & u = silence, s \in J, s' \in J \\ Rel(u|b_P) & u \neq silence, s \neq end, s' = end \\ 0 & o.w. \end{cases} \tag{6.15}$$

6.4.4 Simulation experiment 1

In this simulation, we show how the relevance-based communication model is capable of decreasing the amount of communication needed in a navigation task. We design a map with four possible locations $L_R = \{(0, 6), (1, 1), (5, 1), (5, 4)\}$ (the gray and black tiles from Section 6.4.4) for the reward l^*_R (the black tile from Section 6.4.4). Three walls blocking

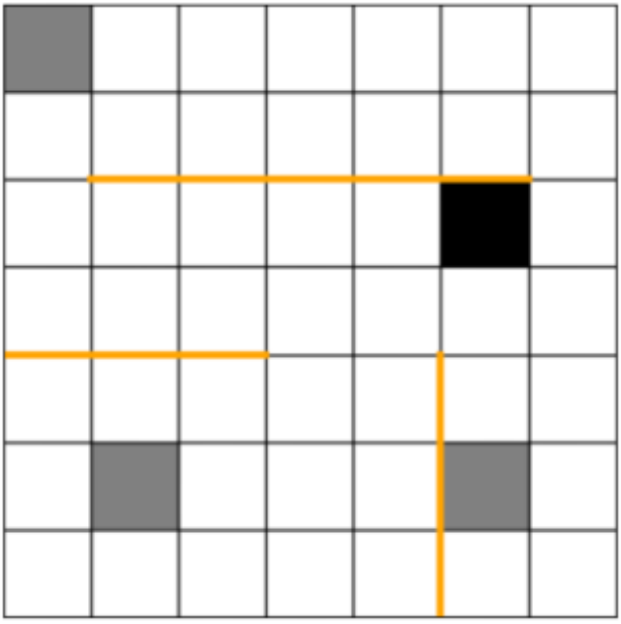

Figure 6.1: **A Sample Map used in Simulation**: Orange lines are walls cannot be surpassed by player; Black tile is the true location of the reward; Gray tiles are other possible locations of the reward than the true one.

agents from passing through are in the map, represented by the orange lines between the tiles. In each game, the player navigates the map to obtain the reward. He can move in four directions: up, right, down, and left from the current tile. After taking the action, he will end up in the next tile in the direction if not blocked by the walls or the map border. If he is blocked by a wall or a map border, he will not move, remaining in the current tile. If the player reaches any of the locations in L_R, the game ends. If he reaches the true reward location l_R^*, the player gains 100 points of reward; if he ends the game in the locations without the reward, he will lose 100 points of reward.

The helper gets the same reward as the player. She knows the map, the location of the reward l_R^*, the and the player's belief b_P. However, she can only help by communicating once to one of the four directions. We conduct two simulation studies: one simulates the scenario when the sender immediately need to choose whether to point at the current timestep or lose the chance to communicate, the other simulates when the helper can use relevance-based model to save the only chance of pointing to later.

In the first scenario, we derive the optimal policy for the helper when the player's belief is $\{(0,6):0.217,(1,1):0.217,(5,1):0.35,(5,4):0.217\}$. The helper can only help the player by pointing to one direction in $\{up, right, down, left\}$ in the player's start, to simulate the one-step decision of communication in Proposition 1. The helper's communicative signal is also constrained by its truthfulness: If the goal is on the top-right side of the player, then the helper can only point to up or right or keep silent. We derive the optimal policy for each, shown in Section 6.4.4.

An illustration of the results is shown in Section 6.4.4. In each graph, the tile in black represents the real location of the of the reward, for example, in the second graph from the left, the reward is in (1, 1). A blue triangle on a tile shows that the helper should point when the player starts from this tile. The arrows in the tile show the direction of the pointing. For example, when the reward is in (1, 1), if the player starts in (3, 2), the sender should either point downwards or leftwards. In contrast, if the player starts in (0, 1), the helper

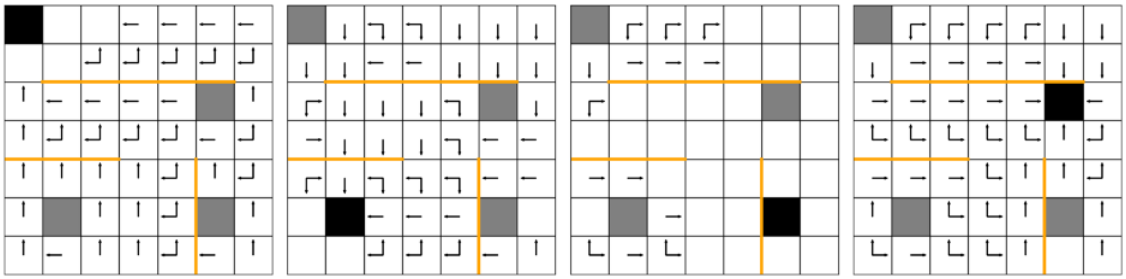

Figure 6.2: **The Helper's Signal Strategy When Pointing is Only Available Immediately**: The arrows represent the direction of the pointing if the player is in this state. If no arrow is presented in this state, the optimal policy for the helper is to stay silent. If the player's belief is correct (c), the helper can communicate less; otherwise, the helper only choose not to communicate when the player is close to the true reward, or when the player cannot help effectively, like in (b) and (d) near (5, 1).

does not need to point because the player is very likely to go to the true reward directly; if the player starts from (6, 1), the helper does not point either. At that state, a pointing to the left may will eliminate the nearest goal (5, 1) from the possible goals, but the player will most likely go to the second nearest location (5, 4) and still fail to reach the true reward. In these cases, either the task is too easy or too difficult to help through communication.

The second simulation experiment is the same as the first simulation experiment except that the sender can use relevance iteration to decide whether to save the one chance to communicate to later. The result is shown in Fig. 6.3. For example, in (d), when the player starts from (3, 2), originally the helper should point to either up or right because the player was going to (5, 4). However, with the belief-space MDP in Proposition 2 to derive the optimal pointing timing, it can keep silent at (3, 2). Because the player was going to (5, 4), the helper can wait until the player goes to (5, 3) then point up. With the belief-space MDP, the helper can save the one pointing signal to when it can help most effectively.

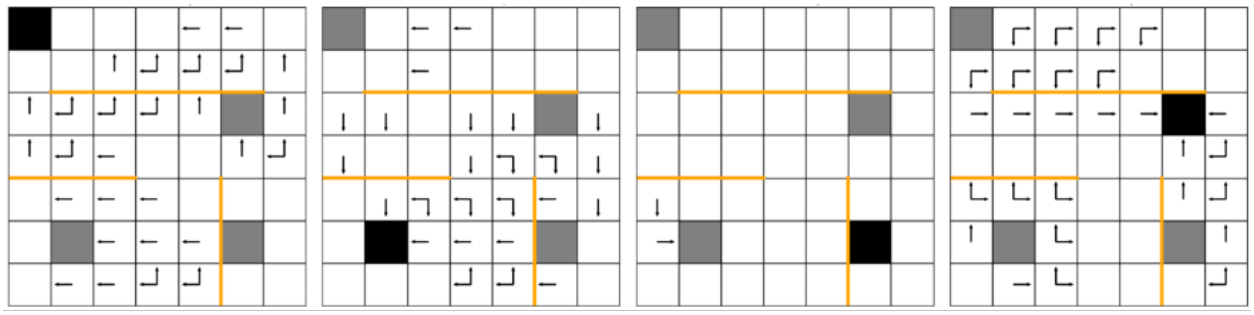

Figure 6.3: **The Helper's Signal Strategy When Pointing Can Be Saved with Relevance Iteration**: Much less communication is observed as it can be saved to the best timing.

6.4.5 Best Timings to Communicate Multiple Times

We introduced the relevance iteration on trajectory-based rMDP to solve the problem of when to communicate when communication is limited to once. In the formulation, since the change of the player's belief depends on not only its current location, but also the history, we changed the state space for the rMDP from physical states L_P in the environment to trajectories J. This change is valid because our constraint that the sender can only communicate once. Her communication trajectory is an array of silence. If we want to extend to scenarios with more changes to communicate, we also need to incorporate the trajectory of communication signals for each time step $j_c = (c^1, c^2, \cdots, c^t)$.

Since the trajectory with communication signals incorporated only changes the process of how belief changes upon receiving the signal, we only need to change the state space of rMDP to the trajectory space $J_{P,C} = \{(j, j_c) | j \in L_P^t, j_c \in U^t, t \in \mathbf{N}\}$. The components of rMDP and sender's MDP can be derived by replacing J in Section 6.4.3 with $J_{P,C}$.

In this section, we will focus on extending the problem in Section 6.4.2 to a problem where the sender can communicate multiple times. In the problem in Section 6.4.2, the player uses an rMDP (L_P, A, T, R_{b_P}). We assume that the player's belief $b_P = P_{L_P,L_R}(l_P, l_R) = P_{L_R}(l_R) I(L_P = l_P)$ only changes depending on the player's current location. The sender

chooses her only chance to communicate with an MDP (S_H, C, T_H, R_H).

Now we extend to the problem where the sender can communicate twice. In this problem, the sender has two phases in selecting actions: The first phase is from the beginning to the sender sending out the first communicative signal; The second phase is from the sender sending out the first communicative signal to the sender sending out the second communicative signal. In the first phase, the sender has two chances to communicate while in the second phase, the sender has one chance to communicate.

$$U_{b_H}(b'_P) = \sum_a \pi(a|b'_P) \left(\sum_{l'_P} \left[(R_{b_H}(l_P, a, l'_P) + V^{(1)}_{b'_{s'}}(s')) T(l_P, a, l'_P) \right] \right) \tag{6.16}$$

where $s' = (l'_P, l_R)$ or end and $b'_{s'}$ is the belief same as b', but the player's location is l'_P instead of l_P. $b'_{s'} = P'_{L_P, L_R}(l'_P, l_R) = P'_{L_R}(l_R) I(L_P = l'_P)$. In Eq. (6.16), $R_{b_H}(l_P, a, l'_P) + V^{(1)}_{b'_{s'}}(s')$ is the maximum utility that the sender gets from the player taking an action a and ending up in l'_P. Then this maximum utility is weighted by the transition probability to l'_P: $T(l_P, a, l'_P)$ and probability of the player taking the action a to calculate the expectation.

With $U_{b_H}(b'_P)$, the sender can calculate the relevance of each communicative signal when two communication opportunities are available.

$$Rel^{(2)}(u|b_P) = \sum_{b'_P} U_{b_H}(b'_P) P(b'_P|b_P, u). \tag{6.17}$$

If the sender chooses not to communicate, she expects to get the reward

$$\sum_a R_{b_H}(l_P, a, l'_P) \pi(a|b_P) T(l_P, a, l'_P). \tag{6.18}$$

Therefore, the reward function for two-chance-for-sender communication is:

$$R^{(2)}_H(s, u, s') = \begin{cases} \sum_a R_{b_H}(l_P, a, l'_P) \pi(a|b_P) T(l_P, a, l'_P) & u = silence, s \in L_P, s' \in L_P \\ Rel^{(2)}(u|b_P) & u \neq silence, s \neq end, s' = end \\ 0 & o.w. \end{cases} \tag{6.19}$$

The transition function for two-chance-for-sender communication is the same as that for one-chance-for-sender communication because as soon as the sender sends out the signal, the two-chance-for-sender phase ends.

$$T_H^{(2)}(s,u,s') = \begin{cases} \sum_a \pi(a|b_P)T(l_P,a,l'_P) & u = silence, s \in L_P, s' \in L_P \\ 1 & u \neq silence, s \neq end, s' = end \\ 0 & o.w. \end{cases} \tag{6.20}$$

With the MDP components $(S_H, C, T_H^{(n)}, R_H^{(n)})$, we can solve the MDP to obtain the optimal communication strategy at each timestep as well as the value function $V_{b_P}^{(n)}(s), s \in S_H$. Then we can use them to derive $(S_H, C, T_H^{(n+1)}, R_H^{(n+1)})$ and $V_{b_P}^{(n+1)}(s), s \in S_H$, and so on, so that we can derive the best timing to communicate in any n-chance-for-communication problems.

6.5 Deciding Best Timing for Communication Is a POMDP

In Chapter 6.3, we proved that if the sender knows the receiver's belief, then her communication process can be modeled as an MDP. In this section, we generalize the argument: If the sender does not know the receiver's belief, then her communication process can be modeled as a POMDP.

Proposition 3. Assume that the sender knows the state, the receiver's belief, policy, and belief transition after receiving each communicative signal. If the sender can only communicate at one timestep, its communication process can be represented as a POMDP for the sender.

Proof of Proposition 3. The sender H knows the state s, and a set of signals U available. If the receiver P has a POMDP represented by (S, A, Ω, T, R, O) and a policy $\pi(a|b)$ for each possible belief b, then the sender can evaluate the utility of every possible belief $b_P \in \Delta_S$

owned by the receiver. The beliefs possibly owned by the receiver are in a set $B_P \subseteq \Delta_S$. The sender maintains a belief over all the possible beliefs of the receiver, represented as $b_S \in \Delta_{B_P}$. As in Proposition 1, for each belief of the receiver $b_P \in \Delta_S$, the decision-making process for the sender for each b_P is an MDP defined as (Δ_S, C, T_b, R_b). Then the decision-making process for the sender for a set of possible b_P is a POMDP $(B_P, C, \Omega_b, T_b, R_b, O_b)$. Here $B_P \subseteq Delta_S$ is the set of all possible beliefs of the receiver over the world state, $C = U \cup \{silence\}$ is the set of communicative signals available, T_b is the transition function between the beliefs given each communicative signal, R_b is the reward function received in the belief transition. Ω_b is a set of observations that helps the sender estimate the receiver's belief. Examples are the states or trajectories of the receiver. The receiver's actions $a \in A$ or trajectory can also serve as observations in Ω_b if they are perceivable by the sender. The receiver's observations $o \in \Omega$ should not be included in Ω_b even if they can be perceived by the sender because these observations are generated from the physical states instead of the receiver's belief states. The observation function $O_b(b', u, o) = P(o|b', u)$ involves the decision-making process of the receiver: the probability that the sender observes the receiver take action or trajectory if the receiver's belief is b' as derived in Eq. (6.5).

As in Proposition 2, the sender also has access to how the receiver changes his mind according to signals, represented as $T_b : \Delta_S \times C \to \Delta_{\Delta_S} : b, u, b' \mapsto$

$$[I(b' = end)I(u \neq silence) + P(b'|b, u)I(u = silence)]\, I(b \neq end) + I(b = end, b' = end).$$

The reward for the belief transition can be calculated as $R_b : \Delta_S \times C \times \Delta_S \to \mathbf{R} : b, u, b' \mapsto [V_{H,0}(b)I(u \neq silence) + r(s)]\, I(b \neq end)$, since the receiver follows a single-agent POMDP after the communication. $c(u)$ represents the cost of sending the signal u.

6.6 Direction Guide Continued with Goal Inference

In the Direction guide game in Chapter 6.4, we assume that the helper knows the belief of the player from the beginning. However, this is not always true in real life. In this section, we

assume that the helper has an initial goal, but the helper does not know it in the beginning of the game. Instead, the helper starts with a flat prior over all possible goals. Through the gameplay, the player moves towards the goal, from which the helper infers the player's goal. In this game, the helper can only communicate once to the helper by pointing to a direction. The helper in the game engages in both inference and planning processes.

In the Direction Guide in Chapter 6.4, we now assume that the player has a goal $l_R \in L_R$. It could be correct if the player's goal is the true reward $l_R = l_R^*$, or wrong if the player's goal is not the true reward. The helper knows the true reward l_R^*, but does not know the player's goal l_R. The helper observes the trajectory j of the player and needs to adopt Bayes' theorem to infer the goals (C. L. Baker, Tenenbaum, & Saxe, 2007).

$$P(l_R|j) = \frac{P(j|l_R)P(l_R)}{\sum_{l'_R \in L_R} P(j|l'_R)P(l'_R)}. \tag{6.21}$$

Intuitively, the helper needs to infer the player's goal and decide whether or not to communicate to help at the same time.

In this case, the player uses a rMDP (L_P, A, T, R_{l_R}), while the helper uses a POMDP (L_R, C, J, T_H, R_H, O). The helper's state space is the set of all possible reward structures of the player because we assume that the player already has a goal l_R; the helper's action space is the set of communicative utterances available, including silence; the helper's observation space is the set of movement trajectories J of the receiver; the helper's transition function models how the reward structure changes as the helper takes her actions: to communicate or to stay silent. Here we assume that the player's reward structure does not change when the helper stays silent, and changes based on different utterances sent out by the helper with probability $P(b'|b, u)$, but the game for the helper ends with the communication.

$$T_H(l_R, u, l'_R) = \begin{cases} 1 & u = silence, l'_R = l_R, s' \in L_P \\ 1 & u \neq silence, l'_R \neq end, l'_R = end \\ 0 & o.w. \end{cases} \tag{6.22}$$

The helper's reward function R_H is $U_{b_H}(b')$ if the helper communicates and the game ends, or

$$\sum_{l_R} \sum_{l'_P} \pi(a|l_P, l_R) P(l_P, a, l'_P) R(l_P, a|l^*_R) \tag{6.23}$$

if the helper chooses to keep silent while evaluating the player's current move. Since the silence does not change the player's reward structure and after communication, the game ends, the observation function is

$$O_H(l'_R, u, j) = \sum_{t=1}^{T} \sum_{a_t} P(s_{t+1}|a_t, s_t, j) P(a_t|s_t, l'_R). \tag{6.24}$$

By solving the helper's POMDP (L_R, C, J, T_H, R_H, O), we can determine the best timing for the helper to communicate to help the player.

6.6.1 Simulation experiment 2

In this simulation, we show how the relevance-based communication model is capable of inferring the goal of the player and communicating at the best timing. We compare the relevance-based model with two models based on heuristics in communication.

We use the similar maps as shown in Section 6.4.4. There are four possible locations $L_R = \{(0, 6), (1, 1), (5, 1), (5, 4)\}$ (the gray and black tiles) for the reward l^*_R (the black tile). In this simulation, we assume that the player has one preset goal $l_{R,P}$ from the four possible locations. We represent its belief with $b_{l_{R,P}}(l_R) = 0.97I(l_R = l_{R,P}) + 0.01I(l_R \neq l_{R,P})$Without any communication, the player will move to this goal with an optimal policy solved by value iteration for MDP $(L_P, A, T, R_{l_{R,P}})$. However, the player's goal is not known to the helper, and needed to be inferred by the player with Eq. (6.21).

The player navigates the maps with a single-agent POMDP model. Each step of navigation costs 5 points, even if the player does not move in this step. If the player reaches any of the goal, the game ends. If the player steps in the true goal, he will receive 100 points of reward; if the player steps in a wrong goal, he will receive -100 points of reward.

The helper knows the true goal and can see the player navigating the map. However, she does not know the player's belief of where the true goal is. She can only communicate with the player once with an utterance selected from the 4 directions: *up*, *down*, *left*, or *right*. Sending out a communicative signal costs 2 points of reward.

We use three models in the simulation: **relevance**-based communication, a heuristic-based communication model **in-line**, and another heuristic-based communication model **distance-increase**. Here we introduce the three models.

In the relevance-based communication model, the helper solves a POMDP (L_R, C, J, T_H, R_H, O). The state space is all possible reward locations L_R. The belief is a distribution over all reward locations, which is the distribution over all the possible beliefs owned by the player. The action space is all the communicative signals including silence. The observations are the trajectories of the player $j_t = (l_P^0, l_P^1, \cdots, l_P^t)$. The transition function is: if the helper does not communicate, the player's belief does not change; if the helper points to one direction, the game ends. The reward function is: if the helper does not communicate, she will receive the reward $\sum_a R_{b_H}(l_P, a, l'_P)\pi(a|b_P)T(l_P, a, l'_P)$, otherwise she will receive $Rel(u|b_P) = U_{b_H}(b^u)$. The observation function is given each possible player's goal, if the helper does not communicate, she will observe the trajectory with probability $P(j|l_R) = \prod_{t=0} \sum_a \pi(a|l_P^t, l_R)P(l_P^{t+1}|l_P^t, a)$; if the helper communicates, the game ends with no more observations provided.

In the in-line model, the helper will point to the direction whenever the player is in line (in the same row or column) with the true reward goal. In the distance-increase model, the helper will point to the direction of the true goal if the Manhattan distance between the player's location and the true goal increases. The players will also interpret the signal correspondingly.

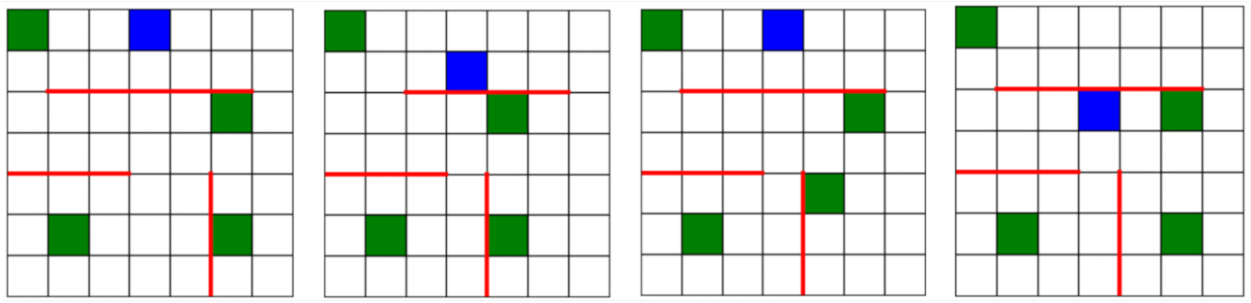

Figure 6.4: **Maps Used in Simulation** Blue tiles represent the starting points of the player, red lines represent walls, green tiles represent possible goals. From the left to right, (a) balanced map, (b) map designed for goal inference + relevance model, (c) map designed for in-line model, and (d) map designed for distance-increase model

6.6.1.1 Maps

We use 4 maps in our simulation, as shown in Fig. 6.4.

In each map, the blue tile represents the starting point of the player. The green tiles represents the possible goals. The red lines represent the walls. If the player attempts to travel against the border of the map or a wall, he will not move.

We design Map 2 for the relevance-based communication model, because it is easier to infer the player's goals in this map. We design Map 3 for the in-line model, because in this map, no possible goals are in the same row or the same column, making the pointing signals less overloaded. We design Map 4 for the distance-increase model because in this map, the player starts from the middle of the map, making it easier for the player to increase his distance from each goal. Map 1 is a more general map that is not designed for a specific model.

6.6.1.2 Conditions and Procedure

We run the 10 simulations for each of the 3 models on each of the 4 maps, with the player's goal being each of the 4 possible goals and the true goal being each of the 4 possible goals.

Model	Map 1	Map 2	Map 3	Map 4
Relevance	**22.88**	**23.88**	16.67	**70.69**
In Line	-12.18	12.33	**19.49**	-3.18
Distance Increase	-6.16	-13.89	8.09	9.91

Table 6.1: The Average Reward Gained Using Each Model on Each Map

In each trial, the player starts from the begining and moves to its own goal. Moving each step costs 5 points. The helper uses her own model to point to one of the four directions at some point, but she can only point once. Pointing costs 2 points. After she points, she cannot communicate anymore. If the player reaches the true goal in the map, he will receive 100 points. If the player reaches a wrong goal in the map, he will lose 100 points

We recorded the reward gained by the player in each of the 1920 simulation trials.

6.6.1.3 Results and Discussion

The average reward gained by the player using each model on each map is shown in Table 6.1.

The pairs of player and helper using relevance-based communication model achieve the highest average reward in 3 out of the 4 maps. The only map that it places the second is Map 3, the map that we designed specific for the in-line model. However, for this map, the performance of the relevance model is still high, close to the performance of the in-line model.

In Map 4, the relevance-based model receives a much higher average reward than the other two models, including the distance-increase model for which the map is designed. We conjecture that this is because when the player starts from the middle, it is also easy for us to infer his goal. When the sender can infer the player's original goal easily, it will be much easier for her to send out a signal that can help.

CHAPTER 7

General Discussion

In this dissertation, I proposed a relevance-based model for human communication. I studied how relevance can be successfully applied to explain human overloaded, indirect, and sparse communication with derivation from utility theory and decision-making theory, and with inspiration from information theory, cognitive science, and developmental psychology. In this chapter I discuss a few topics that arise from studying the relevance-based model.

7.1 How Relevance Enables Sparse, Overloaded, and Indirect Human Communication

In the dissertation, we use the parameter relevance to capture the sparse, overloaded, and indirect nature of human communication. Communication is a (part of a) joint task, in which the jointness constructs a large set of common knowledge, including the joint goal, each other's perception of the world, and the commitment to help. Sending out the communicative signal is treated as an action of the sender. The effect of this signal is to change the receiver's mind so that the joint task can be accomplished with maximum the utility. In language games like reference games, the joint task is for the receiver to understand the sender's language. In physical tasks like cooking (Wu et al., 2021) or crafting (Gong et al., 2023), we propose that communication helps in a more measurable way: increasing the already defined utility. In both scenarios, the sender considers the receiver's mind change and calculates the utility after receiving each signal, which we call the relevance of the signal.

With the joint task as the context, humans are capable of using overloaded and indirect signals. Once a signal is received, no matter it is as long as a dissertation or as short as a simple *Ahh*, the receiver initiates the inference process. In the inference, the receiver reasons about what this signal represents, what else could have been said, and why the signal is sent out. In Chapter 4, we define the parameter relevance to summarize this reasoning process by taking into account the joint task, the current status in accomplishing the task, and the cooperative motivation assumption. With relevance, the receiver will be more likely to adopt the interpretation that helps the joint task the most, which solves the problem of deciphering overloaded communicative signals with or even without a literal meaning.

The reasoning process can depend on the literal meaning of the signal, but more importantly the joint task. In joint physical tasks, the joint goal may be independent of the literal meaning of the signal, which results in the indirect nature of communication. For example, when the goal is a mom to urge her kids to go to bed. She does not need to point to the bed whose literal meaning aligns with the goal. When she points to the clock whose literal meaning is independent from the goal bed, the kids can make the connection by inferring with the joint goal. The mom can also simply call the kids' names. In Chapter 4 and 5, we applied relevance-based communication in the Guided Wumpus Hunt game. In Chapter 4, the communicative signals stench, glitter, and silence do not reveal more information about the reward structure than the single observation but can help much better with the inference. In Chapter 5, the signals of pointing to each Wumpus have the same literal meaning, but their effect to help the joint goal is different, resulting in human participants' preferences between the signals. T. Sumers et al. (2022) linearly combines truthfulness (the literal meaning) and relevance (the utility) in communication to model this joint inference.

In Chapter 6, we discussed the sparsity of communicative signals, which may result from the effect of communicative signals: to change the mind. In Chapter 4, we defined the relevance between two minds to depict the sender's evaluation of the status of accomplishing the task. The sender should send out the signal when it helps the utility of the joint task

to the largest extent. After the mind is changed, the relevance of between two minds is low, so the sender does not need to send another communicative signal. Otherwise, it may jeopardize the on-track joint task.

It is worth noting that relevance is not symmetric between the sender and receiver. Instead, it is evaluated from the sender's perspective because sending out communicative signal is the sender's action. In Chapter 3, we introduced the paternalistic evaluation of the receiver's action, which provides a common knowledge of how to cross the two minds in communication. Both the sender and the receiver evaluate the relevance in a paternalistic manner and know the other party does it too. This process demands strong ToM capacity, which can be found in pointing chimpanzees and human infants much earlier than the development of language.

7.2 Do Humans Use Complex Models Like Relevance-Based Communication

In Chapter 5, we successfully predicted human participants' choices with our relevance-based model. Also, agents that communicate with the relevance-based model are more well-received in human participants than agents that communicate with a heuristic-based model. However, it remains a question: do humans actually calculate relevance in real life communication?

Relevance is difficult to calculate. It needs strong planning capacity to evaluate and predict (joint) actions in partially observable environments over time, ToM capacity to take and coordinate other people's perspective, as well as the capacity to recursively reason about mind. When designing an artificial agent with the relevance model, computation is complex in a) planning in the environment with POMDP solvers, b) estimating the change of receiver's mind after receiving the signal to calculate relevance, and c) recursive inference. However, in practice, humans usually decide what to communicate and how to understand communication in a blink of time. The stark contrast exposes a gap: do humans actually use relevance in

communication?

7.2.1 Planning Capacity

Humans have good planning capacity that has been intensively studied in computer science and reinforcement learning (Sutton & Barto, 2018). Specifically oriented to partially observable sequential decision-making problems, multiple POMDP solvers were proposed (Pineau et al., 2003; Spaan & Vlassis, 2005; Kurniawati, Hsu, & Lee, 2009; Silver & Veness, 2010). Most POMDP solvers struggle with the curse of dimensionality and the curse of history. When the environment is large with a large number of possible states or the time horizon of planning is high, the computational complexity of POMDP solvers will increase more than exponentially. However, human planning does not suffer from these curses too much. First, we seem to use dimension reduction mechanisms and high-level actions (Sutton, Precup, & Singh, 1999) to largely reduce the cardinality of state space and action space, and the planning horizon in problem solving. For example, when we plan to take a cup of coffee, we only need to plan with a high-level action go to the kitchen instead of planning each footstep we take. Second, when we plan, we do not consider too far horizon in the future. We only need to think of what action to take in the current time step. Luckily, a lot of policy trees share the same initial root action. We only need a partition of the belief space and assign a first action for each. For example, if we do not know which shop in the food court sells the new taco, we do not need to take different actions based on our knowledge when we are still in the classroom. No matter what our belief is, we need to go to the food court first anyway. Third, for the same reason, we can give up some accuracy in planning in exchange for speed. A lot of POMDP solvers are anytime solvers, which means they can stop updating their value function after any number of iterations. The more iterations that they have, the more accurate the estimate is for the value function, but more time is used. Since we do not need precise planning, we can stop the planning very soon as long as we have a good estimate of the root action of the policy trees. Fourth, thanks to cooperative motivation occupied by

humans, when we do not have a great plan, we can always communicate with other people so that they can help us, teach us, and cooperate with us. It is also likely that humans use heuristics in planning, but heuristics may be an appearance of the strategies above.

7.2.2 Belief estimation and relevance calculation

With ToM, we maintain a representation of how the mind works: how components like belief, desire and intention interact with each other and how they determine our actions. We can also reverse this process and infer the mind components of others from observing their actions. In our relevance-based communication model, we assume that the sender has an accurate estimate of the receiver's belief, while the receiver infers the sender's belief from the sender's communicative actions. This may not be the case in the real world. When humans infer others' belief from their actions, it is common to know what they know and what they do not know. However, it is not common to say that the belief of another individual is 0.23 in this state and 0.87 in the other state. Does it mean we should not model relevance calculation with precise belief change?

First, empirical study shows that humans maintains an implicit probabilistic representation of belief, like in goal inference (C. L. Baker et al., 2007). Second, from a modeling perspective, belief estimation and relevance calculation do not need to be precise. Belief estimation is used for policy prediction. Due to the continuity of the value function of POMDPs, two beliefs that are close should not have large difference in their values on the receiver's single-agent POMDP value function $V(b_P)$, hence in their predicted policies. This is true especially when the planning horizon is low for which the value function is the maximum of a few linear functions on the belief space. Therefore, as long as we have a rough estimate of the receiver's belief, we can predict its action.

Similar to action prediction, the action evaluation in relevance calculation does not require accurate belief estimate either. As demonstrated in Chapter 4, we demonstrate that the receiver's belief evaluated by the sender value function $U_{b_H}(b_P)$ is the intercept of the tangent

plane at the receiver's belief on the sender's belief. It is no longer continuous on the receiver's belief space. However, in a low horizon POMDP, the value function is the maximum of a few linear functions, so the $U_{b_H}(b_P)$ will not change too much for two b_P close to each other. Therefore, the relevance calculation is not affected too much. More importantly, relevance is a relative measurement, whose rank is more important than its actual value. The receiver only needs to infer what sender's mind is the most relevant. With the estimation of how the signals change the receiver's minds, the sender only needs to consider which signal is the most relevant. Therefore, in practice, a rough estimation of relevance is sufficient as long as the rank is maintained. The minds and signals with lower relevance do not need to be accurate in estimation either. Therefore, we can save calculation in relevance.

Third, even we can tolerate minor bias in belief estimation, relevance calculation, and belief change simulation, they are still difficult in human life. Estimation error happens every day. However, human senders do not have to change the receiver's mind to ideal with only one communicative signal. When the sender sends out a signal to a stranger receiver, it is very likely that the sender cannot accurately estimate how it changes the receiver's belief. However, the sender can observe the receiver's action to infer its belief after receiving the signal. Then more communicative signals can be sent out to make up in the subsequent timesteps until the receiver's belief is changed to ideal. AI researchers prompt large language models in this chain-of-thought manner, largely improving the problem-solving capacity of large language models. Through this process, the sender and the receiver can learn the model of the effects of each signal and become more in tune with your cooperators.

7.2.3 Recursion

Recursion makes computation much more complex. In the relevance-based communication model as well as most communication models like RSA (Frank & Goodman, 2012) and CI-POMDP (P. Gmytrasiewicz, 2020), the sender simulates the belief change of the receiver while the receiver tries to infer this simulation process. As a consequence, the computational

complexity is much higher than single-agent planning and inference. Take the sender in relevance-based communication as an example. The sender needs to evaluate the entire belief space of the receiver to compute the relevance. However, because of the recursion, the evaluation of each belief point in the receiver's belief space needs the simulation on the entire belief space of the sender. This process puts evaluation in two continuous spaces instead of one single belief point in a single-agent planning, largely complexifying the computation. Similarly in CI-POMDP, each agent maintains a model of the other agents including their POMDP components and type which measures how cooperative and competitive they are. The recursion continues until no knowledge can be utilized to assume a non-flat belief, for example, if you cannot see me. The computation in i-POMDP and CI-POMDP is very complex, nearly impossible for human to calculate with mind. Should we keep using recursion in our computational models?

Computationally, we do not need a high level of recursion in our relevance model. In Chapter 4, the relevance-based agents achieve a near optimal performance even with only one level of recursion. As long as agents are taking the mind of the other agent into consideration in their planning, the performance is high enough. Therefore, one level of recursion may be enough for human communication. Higher levels of recursion could be ignored as the computation complexity outweighs the gain in performance.

Theoretically, human communication is highly transparent. The context of the joint task, the world, and the cooperative assumption is commonly known to each agent, so they do not need to recursively reason about the context. With common knowledge, most of the components in the POMDP does not need to be recursively reasoned in CI-POMDP. In Chapter 2, we show that when agents communicate, the messages are not the same as individual observations. Instead, they should be put into common knowledge so that it does not engage in high level recursions. In addition, in the joint task with communication, agents may shift from their single-agent utility maximizing action pattern to an action pattern that clearly show their intention to facilitate potential relevance-based communication, as we

conjectured in Chapter 6.

In all, we do use recursive reasoning in real life, so do we in models. However, we do not need to worry about the computational complexity brought about by high-level recursions. A low-level recursion is sufficient for a high performance in the joint task, meanwhile the transparent common knowledge and decision-making under the jointness assumption help reduce recursion.

7.3 What Can We Learn from Relevance-based Communication in Designing Artificial Intelligence Systems?

Our relevance-based communication model roots deep in visual communication, opposed to language communication, the hot topic of current AI trend. In this framework, agents are modeled with reward-driven mind-based planning capacity, ToM, and a cooperative motivation. Most current question-answering AI models insinuate a cooperative assumption because answering questions is to finish a joint task together. However, a more transparent common knowledge that benefits from the cooperative assumption could be a challenge for a lot of language models, so a common problem is that their answers are much longer than human verbal responses. I hope my research can inspire more researches on modeling visually-grounded human communication, which includes establishing a trangulation in joint attention, forming a shared commitment to a task (Tomasello, Carpenter, Call, Behne, & Moll, 2005; Tang, Stacy, Zhao, Marquez, & Gao, 2020; Tang et al., 2022). Indeed, theories suggest that the primary objective of conversation is to manipulate each other's attention, thereby fostering a shared common ground of information among cooperative agents (Stacy, Zhao, Zhao, Kleiman-Weiner, & Gao, 2020; Stacy et al., 2021).

More researchers start to study ToM in artificial intelligence. Large language models have been shown to have ToM ability (Sap et al., 2022). However, they are more likely the study of ToM in question-answering models is limited to probing the model with classical

tasks of ToM in developmental psychology. ToM has also been applied to computer vision studies (Zhao et al., 2015; Gao et al., 2020). We hope that ToM-based model of agent can thrive in modeling human-AI interaction and designing reinforcement learning systems with human feedback. With ToM, we are more likely to create an AI system that are well-received by humans, and learn more implicitly than only from human languages.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on machine learning* (p. 1).

Arechavaleta, G., Laumond, J.-P., Hicheur, H., & Berthoz, A. (2008). An optimality principle governing human walking. *IEEE Transactions on Robotics*, *24*(1), 5–14.

Baker, C., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2007). Goal inference as inverse planning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29).

Barnes-Holmes, Y., McHugh, L., & Barnes-Holmes, D. (2004). Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today*, *5*(1), 15.

Barth, J., & Call, J. (2006). Tracking the displacement of objects: a series of tasks with great apes (pan troglodytes, pan paniscus, gorilla gorilla, and pongo pygmaeus) and young children (homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(3), 239.

Benotti, L., & Blackburn, P. (2011). Classical planning and causal implicatures. In *International and interdisciplinary conference on modeling and using context* (pp. 26–39).

Bobu, A., Scobee, D. R. R., Fisac, J. F., Sastry, S. S., & Dragan, A. D. (2020). Less is more: Rethinking probabilistic models of human behavior. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (p. 429–437). doi: 10.1145/3319502.3374811

Bratman, M. (1987). *Intention, plans, and practical reason.* Harvard University Press.

Brinck, I. (2004). The pragmatics of imperative and declarative pointing. *Cognitive Science Quarterly*, *3*(4), 429–446.

Buchsbaum, D., Bridgers, S., Skolnick Weisberg, D., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2202–2212.

Bullinger, A. F., Wyman, E., Melis, A. P., & Tomasello, M. (2011). Coordination of chimpanzees (pan troglodytes) in a stag hunt game. *International Journal of Primatology*, *32*, 1296–1310.

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PloS one*, *12*(4), e0173793.

Buttelmann, D., Carpenter, M., Call, J., & Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental science*, *10*(4), F31–F38.

Butterworth, G., Simion, F., et al. (2013). *The development of sensory, motor and cognitive capacities in early infancy: From sensation to cognition*. Routledge.

Campbell, J. (2005). Joint attention and common knowledge. *Joint attention: Communication and other minds*, 287–297.

Carpenter, M., & Liebal, K. (2011). Joint attention, communication, and knowing together in infancy. In *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience* (pp. 159–181). MIT Press.

Carpenter, M., Tomasello, M., & Striano, T. (2005). Role reversal imitation and language in typically developing infants and children with autism. *Infancy*, *8*(3), 253–278.

Cassandra, A. R., Kaelbling, L. P., & Kurien, J. A. (1996). Acting under uncertainty: Discrete bayesian models for mobile-robot navigation. In *Proceedings of ieee/rsj international conference on intelligent robots and systems. iros'96* (Vol. 2, pp. 963–972).

Clissa, L., Lassnig, M., & Rinaldi, L. (2023). How big is big data? a comprehensive survey of data production, storage, and streaming in science and industry. *Frontiers in big*

Data, *6*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*, 2493–2537.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, *267*(5199), 843–848.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doshi, P., Gmytrasiewicz, P., & Durfee, E. (2020). Recursively modeling other agents for decision making: A research perspective. *Artificial Intelligence*, *279*, 103202.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Du, Y., Kosoy, E., Dayan, A., Rufova, M., Abbeel, P., & Gopnik, A. (2023). What can ai learn from human exploration? intrinsically-motivated humans and agents in open-world exploration. In *Neurips 2023 workshop: Information-theoretic principles in cognitive systems.*

Eilan, N. (2005). *Joint attention: Communication and other minds: Issues in philosophy and psychology*. Oxford University Press, USA.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.

Fletcher, G. E., Warneken, F., & Tomasello, M. (2012). Differences in cognitive processes underlying the collaborative activities of children and chimpanzees. *Cognitive Development*, *27*(2), 136–153.

Franco, F. (2005). Infant pointing: Harlequin, servant of two masters. In *Joint attention: Communication and other minds* (pp. 129–164). Oxford University Press New York.

Franco, F., & Gagliano, A. (2001). Toddlers' pointing when joint attention is obstructed. *First Language*, *21*(63), 289–321.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gao, X., Gong, R., Zhao, Y., Wang, S., Shu, T., & Zhu, S.-C. (2020). Joint mind modeling for explanation generation in complex human-robot collaborative tasks. In *2020 29th ieee international conference on robot and human interactive communication (ro-man)* (pp. 1119–1126).

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287–292.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gmytrasiewicz, P. (2020). How to do things with words: A bayesian approach. *Journal of Artificial Intelligence Research*, *68*, 753–776.

Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, *24*, 49–79.

Golinkoff, R. M. (1993). When is communication a 'meeting of minds'? *Journal of child language*, *20*(1), 199–207.

Gong, R., Huang, Q., Ma, X., Vo, H., Durante, Z., Noda, Y., ... others (2023). Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. *Developmental psychology*, *45*(5), 1430.

Greenberg, J. R., Hamann, K., Warneken, F., & Tomasello, M. (2010). Chimpanzee helping

in collaborative and noncollaborative contexts. *Animal Behaviour*, *80*(5), 873–880.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). Brill.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, *476*(7360), 328–331.

Hamann, K., Warneken, F., & Tomasello, M. (2012). Children's developing commitments to joint goals. *Child development*, *83*(1), 137–145.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771–785.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal behaviour*, *68*(3), 571–581.

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in neural information processing systems*, *30*.

Hawkins, R. X., Frank, M., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Cogsci*.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, *317*(5843), 1360–1366.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, *150*(11), 2246.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in neural information processing*

systems, *29*.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jiang, K., Stacy, S., Chan, A., Wei, C., Rossano, F., Zhu, Y., & Gao, T. (2021). Individual vs. joint perception: a pragmatic model of pointing as smithian helping. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Jiang, K., Stacy, S., Dahmani, A. L., Jiang, B., Rossano, F., Zhu, Y., & Gao, T. (2022). What is the point? a theory of mind model of relevance. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

John, M., Duguid, S., Tomasello, M., & Melis, A. P. (2019). How chimpanzees (pan troglodytes) share the spoils with collaborators and bystanders. *PloS one*, *14*(9), e0222795.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Kaelbling, L. P., & Lozano-Pérez, T. (2013). Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, *32*(9-10), 1194–1227.

Kao, J., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Karten, S., Tucker, M., Kailas, S., & Sycara, K. (2023). Towards true lossless sparse communication in multi-agent systems. In *2023 ieee international conference on robotics and automation (icra)* (pp. 7191–7197).

Kearns, M., Mansour, Y., & Ng, A. (1999). Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, *12*.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kurniawati, H., Hsu, D., & Lee, W. S. (2009). Sarsop: Efficient point-based pomdp plan-

ning by approximating optimally reachable belief spaces. In O. Brock, J. Trinkle, & F. Ramos (Eds.), *Robotics: Science and systems iv* (p. 65-72). MIT Press.

Lascarides, A., & Stone, M. (2009). A formal semantic analysis of gesture. *Journal of Semantics*, *26*(4), 393–449.

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

Leavens, D. A., & Hopkins, W. D. (1998). Intentional communication by chimpanzees: a cross-sectional study of the use of referential gestures. *Developmental psychology*, *34*(5), 813.

Liebal, K., Behne, T., Carpenter, M., & Tomasello, M. (2009). Infants use shared experience to interpret pointing gestures. *Developmental science*, *12*(2), 264–271.

Liszkowski, U. (2009). Human twelve-month-olds point cooperatively to share interest with and helpfully provide information for a communicative partner. *Gestural communication in nonhuman and human primates*, 123–140.

Liszkowski, U., Carpenter, M., Striano, T., & Tomasello, M. (2006). 12-and 18-month-olds point to provide information for others. *Journal of cognition and development*, *7*(2), 173–187.

Liszkowski, U., Schäfer, M., Carpenter, M., & Tomasello, M. (2009). Prelinguistic infants, but not chimpanzees, communicate about absent entities. *Psychological Science*, *20*(5), 654–660.

Littman, M. L., Cassandra, A. R., & Kaelbling, L. P. (1995). Learning policies for partially observable environments: Scaling up. In *Machine learning proceedings 1995* (pp. 362–370). Elsevier.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Lovejoy, W. S. (1991). Computationally feasible bounds for partially observed markov decision processes. *Operations research*, *39*(1), 162–175.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Martin, A., Lin, K., & Olson, K. R. (2016). What you want versus what's good for you: Paternalistic motivation in children's helping behavior. *Child development*, *87*(6), 1739–1746.

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology*, *120*(2), 154.

Melis, A. P., Schneider, A.-C., & Tomasello, M. (2011). Chimpanzees, pan troglodytes, share food in the same way after collaborative and individual food acquisition. *Animal Behaviour*, *82*(3), 485–493.

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological science*, *27*(12), 1550–1561.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603–613.

Montufar, G., Ghazi-Zahedi, K., & Ay, N. (2015). Geometry and determinism of optimal stationary control in partially observable markov decision processes. *arXiv preprint arXiv:1503.07206*.

Oliehoek, F. A., Amato, C., et al. (2016). *A concise introduction to decentralized pomdps* (Vol. 1). Springer.

OpenAI. (2023). *Gpt-4 technical report.*

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations.*

Pineau, J., Gordon, G., Thrun, S., et al. (2003). Point-based value iteration: An anytime algorithm for pomdps. In *Ijcai* (Vol. 3, pp. 1025–1032).

Reddy, S., Levine, S., & Dragan, A. (2021). Assisted perception: optimizing observations to communicate state. In *Conference on robot learning* (pp. 748–764).

Ross, S., Pineau, J., Paquet, S., & Chaib-Draa, B. (2008). Online planning algorithms for pomdps. *Journal of Artificial Intelligence Research*, *32*, 663–704.

Royka, A., Chen, A., Aboody, R., Huanca, T., & Jara-Ettinger, J. (2022). People infer communicative action through an expectation for efficient communication. *Nature Communications*, *13*(1), 4160.

Royka, A. L., Török, G., & Jara-Ettinger, J. (2023). Guiding inference: Signaling intentions using efficient action. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: a modern approach* (3rd ed.). Prentice Hall.

Sap, M., LeBras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Shani, G., Pineau, J., & Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, *27*, 1–51.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Silver, D., & Veness, J. (2010). Monte-carlo planning in large pomdps. *Advances in neural information processing systems*, *23*.

Smith, A. (2010). *The theory of moral sentiments*. Penguin. (Original work published 1759)

Spaan, M. T., Gordon, G. J., & Vlassis, N. (2006). Decentralized planning under uncertainty for teams of communicating agents. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems* (pp. 249–256).

Spaan, M. T., & Vlassis, N. (2005). Perseus: Randomized point-based value iteration for pomdps. *Journal of Artificial Intelligence Research*, *24*, 195–220.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press Cambridge, MA.

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., Kleiman-Weiner, M., & Gao, T. (2021). Modeling communication to coordinate perspectives in cooperation. *arXiv preprint arXiv:2106.02164*.

Stacy, S., Zhao, Q., Zhao, M., Kleiman-Weiner, M., & Gao, T. (2020). Intuitive signaling through an "*Imagined We*". In *Cogsci.*

Sumers, T., Ho, M. K., Griffiths, T. L., & Hawkins, R. (2022, Oct). *Reconciling truthfulness and relevance as epistemic and decision-theoretic utility.* PsyArXiv. Retrieved from `psyarxiv.com/e9m3j` doi: 10.31234/osf.io/e9m3j

Sumers, T. R., Hawkins, R. D., Ho, M. K., & Griffiths, T. L. (2021). Extending rational models of communication from beliefs to actions. *arXiv preprint arXiv:2105.11950*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, *112*(1-2), 181–211.

Tang, N., Gong, S., Zhao, M., Gu, C., Zhou, J., Shen, M., & Gao, T. (2022). Exploring an imagined "we" in human collective hunting: Joint commitment within shared intentionality. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Tang, N., Stacy, S., Zhao, M., Marquez, G., & Gao, T. (2020). Bootstrapping an imagined we for cooperation. In *Cogsci.*

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

Tomasello, M. (2020). Why don't apes point? In *Roots of human sociality* (pp. 506–524). Routledge.

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child development*, 1067–1080.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–691.

Tomasello, M., Carpenter, M., & Hobson, R. P. (2005). The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child development*, i–152.

Tomasello, M., & Haberl, K. (2003). Understanding attention: 12-and 18-month-olds know what is new for other persons. *Developmental psychology*, *39*(5), 906.

Tong, X., Zhang, Z., Wang, J., Huang, C., & Debbah, M. (2021). Joint multi-user communication and sensing exploiting both signal and environment sparsity. *IEEE Journal of Selected Topics in Signal Processing*, *15*(6), 1409–1422.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Nips*.

Vaish, A., Carpenter, M., & Tomasello, M. (2016). The early emergence of guilt-motivated prosocial behavior. *Child development*, *87*(6), 1772–1782.

Van der Goot, M. H., Tomasello, M., & Liszkowski, U. (2014). Differences in the nonverbal requests of great apes and human infants. *Child Development*, *85*(2), 444–455.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vygotsky, L. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS biology*, *5*(7), e184.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *science*, *311*(5765), 1301–1303.

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, *15*(6), 728–755.

Wilson, D., & Sperber, D. (2006). Relevance theory. *The handbook of pragmatics*, 606–632.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wittgenstein, L., & Anscombe, G. (2001). *Philosophical investigations: the german text, with a revised english translation* (3rd ed.). Blackwell. (Original work published 1953)

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.

Xuan, P., Lesser, V., & Zilberstein, S. (2001). Communication decisions in multi-agent cooperation: Model and experiments. In *Proceedings of the fifth international conference on autonomous agents* (pp. 616–623).

Yamamoto, S., Humle, T., & Tanaka, M. (2012). Chimpanzees' flexible targeted helping based on an understanding of conspecifics' goals. *Proceedings of the National Academy of Sciences*, *109*(9), 3588–3592.

Zhao, Y., Holtzen, S., Gao, T., & Zhu, S.-C. (2015). Represent and infer human theory of mind for human-robot interaction. In *2015 aaai fall symposium series*.