Learning Ownership based on Theory of Mind

Zhihao Cao
Department of Automation
Tsinghua University
caozh20@mails.tsinghua.edu.cn

Zian Wang
Department of Automation
Tsinghua University
wza20@mails.tsinghua.edu.cn

Abstract

Social norms are a significant part of human society. It is impossible for our agents to interact with humans in social environments if they are not able to understand and conform to social norms. Ownership is an essential part of social norms. The concept of territory, company, and property is all based on ownership. Recent works represent and infer ownership using simple inference, or represent it with a probabilistic graph and infer it using predicate-based norms. However, they all link ownership with action rather than intent, ignoring agents' mental states. This paper presents a Theory-of-Mind (ToM) based model capable of simulating agents' change of intention based on ownership. With the given intention, our agents can choose different actions based on the state of the world. By combining ownership with intention rather than directly with action, we can study how agents' mental states will change with the understanding of ownership, which can give our agents a stronger generalization ability.

1 Introduction

Social norms are essential when we are interacting with others socially. As Ferreira et al. [6] have argued, if our agents want to cooperate with humans, they must have a "norm capacity." Ownership is an essential part of them. If our agents can not understand ownership, they will invade our territory, get our property, like money and jewelry, get away with them and sell them for their own goals. If this could happen, our clients would never allow our agents to get into their lives. Then it will be impossible for our agents to cooperate with humans completely. It is also one of the oldest social norms, as it exists in all kinds of societies, including primitive societies in Africa. Ownership has accompanied us since society came into existence. Thus if we can figure out the beginning of ownership and how to represent, study and use it, it will provide a valuable paradigm for researching other social norms, like power and etiquette.

There is a relatively clear definition and description of ownership in law. It is equaled to "productive labor of the owner" for economists [19]. It is the foundation of many other practical concepts, such as money, trade, and theft. The process and mechanics of ownership are pretty complex, as there are many ways that one can gain, transfer or lose ownership of items.

We must point out, however, that when the concept of ownership is described by the product of our language, like the law, it must first appear in human cognition. The language is a system of encoding, and the object to be encoded must have existed before the encoding method encoded it. Scientific studies have shown that the concept of ownership has formed from a very early age. Ellis [5] observed possessive behavior in children at a very early age, even before they can use language to express the concept of ownership. Moreover, new evidence shows that an early understanding of ownership helps infants efficiently organize objects in memory[14]. Thus, to study ownership thoroughly, we must start with nonverbal behavior rather than language.

In our daily life, we often claim an item as our own and get angry when someone else uses it without permission. At the same time, we have a certain sense of ownership of other people, such as wife to husband, and children to parents. Veblen [19] suggested that this is the origin of ownership, and

the ownership of objects is a generalization of people's ownership. He argued that the concept of ownership of people comes from the ownership of enslaved people during the slave society, which is the origin of ownership[19].

There is rarely a unified framework to provide a complete computation process for the formation and use of the concept of ownership. Recent work either represent and infer ownership with simple rules, or represent it with probabilistic graph and infer it with predicate-based norms. And most of them simply link understanding ownership with action, ignoring agents' mental states. Moreover, there is no good data set that has emerged to annotate people's ownership. For research about mental states, it is a big problem. In this circumstance, researchers have to learn the concepts they are interested in without existing data, verify the validity of the model they proposed through human studies, and ultimately constantly change the model based on the feedback to achieve its human-like performance and explanation. Here we follow this pipeline and research ownership. By using the Theory of Mind (ToM) method, we modeled simple agents with nested intents. Based on this, we established a unified computational framework for modeling agents' ownership concepts. Based on this model, we generated some data, conducted human experiments, and verified the model's and data's validity. The experimental results indicate that this data set can characterize, in part, people's learning and use of ownership concepts well. Based on these data, we discuss whether people form ownership concepts for space, objects, or both.

2 Related work

Ownership Veblen [19] argued that the concept of ownership origins from the possession of enslaved people plundered from other tribes during the slave society. His argument hints that ownership starts with the refusion of others' actions to our possession, which provides a theoretical foundation for our model. Thus, In our model, the ownership relation is achieved by preventing others' approaching action to the possession by the owner. Ellis [5] observed possessive behavior in children at a very early age, even before they can use language to express the concept of ownership. Thus, in this paper, we study ownership not by language, but by nonverbal actions. Friedman [8] proposed that our perception of ownership mainly depends on the principle of first possession. The first possession principle assists us in initializing the ownership relation. We resemble the principle with the nearest principle. Stahl et al. [14] showed an early understanding of ownership helped infants efficiently organize objects in memory. Pierce et al. [12] provided a psychological understanding of ownership. They proved the significance of studying ownership. Tan et al. [15] studied ownership in a limited environment based on robots, which can study fundamental relations of ownership and prevent others' invalid actions depending on them. This paper extends their work by linking ownership with intention rather than direct action. It can help our model generalize to a new situation more quickly.

Intention Woodward [22] showed that even at five months, we could detect others' intentions. Anscombe [1] provided a detailed explanation of intention. Fodor [7] and Samuels [13] show us opinions on the mind modularity problem, which gives us some guidance on how to build our models. Zhu et al. [23] provided a brief introduction about intention. Baker et al. [3] represented intention as several different destinations. He assumed that the agent wanted to move somewhere, and the destinations could only be the marked places. Thus, the difference only lies in the destination. Wang et al. [20] paid attention to the objects that agents wanted to get. In his task, the agents needed to communicate with each other and at last got one target respectively, which made him take the intention as different objects. Lee et al. [11] regarded intention and intentional action as equivalent. They claimed that an intention is achieved when a corresponding action is completed. Ullman et al. [18] distinguished individual intentions from social intentions. Individual intentions meant target objects that agents wanted to get, and social intentions meant helping or hindering others from achieving their intentions. Holtzen et al. [10] expanded the space of intentions, but still equaled intentions with target objects. Tomasello [17] argued that we have four different intents: Individual intent, social intent, communicative intent, and referential intent. Woodward [22] also pointed out that different kinds of intents can have different levels. Wei et al. [21] used Human-attention-object (HAO) graph to represent the relationship between different intents.

3 Framework

3.1 Structure

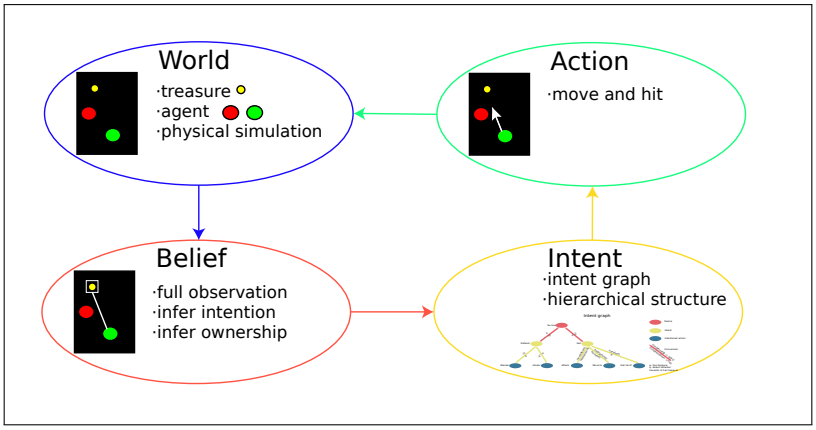

Figure 1: The structure of our model. There are four main parts: world, belief, intention, and action. The world module runs physical simulations based on action and provides information to the belief module. The belief module observes the world, gets information, infers intention and ownership, and provides information to the intention module. The intention module will use the intent graph and the information from the belief module to generate new intentions. The action module will choose an action based on generated intention and the information from the belief module, and then submit the action to the world module. Then a round is completed.

The whole structure is shown in 1. It follows classical Our model has four main parts: World, Belief, Intention, and Action.

World The world consists of agents and treasures, and provides a physical engine to run simulations. For simplification, it is 2-D.

Belief Belief includes belief about the world's physical states and other agents' mental states. We adopt comprehensive observation rather than partial observation for simplification. As the center of our model is to study the change of mental states with the understanding of ownership, the simplification of observation is acceptable. Based on observation, our agents will infer others' intentions. Here we achieve the inference by finding the target others are paying attention to and combining it with the understanding of ownership. We assume that all the agents have the same behavior model, and then the complete observation makes all the agents' inferences about others' intentions the same. Thus, the world can achieve inference instead of the agents in practice.

Intention We use a hierarchical structure to build the link between different intentions. Our agents will choose their intentions based on their beliefs, current intentions, and the intention graph.

Action Action is the available act that our agents can choose to achieve their intention. It will be chosen depending on the agents' intentions.

3.2 Intention

3.2.1 Intent graph

As mentioned in related work, we usually represent intention in an atomic way. However, Tomasello [17] argued that we have four different intents: Individual intent, social intent, communicative intent, and referential intent. Tomasello's argument hints at one possibility: the intention can have a very complex structure. Instead of atomic representation or only four layers representation, there is still a more complex structure about intention. Thus we choose to represent the intention with a graph, to simulate any possible structure. The meaning of each part is:

Node Nodes represent different kinds of intention. The intention can be pure intent, normal intention, or intentional action. Not every intention can correspond to a specific action policy. Only the intentional action node, or called leaf intention node, corresponds to a specific action policy and is able to affect the real world.

Direction Direction represents the conversion between different intentions. The previous intention generated by the intention graph is called the father intention, and the intention generated by this intention is called the child intention. The father-children relationship is relative rather than constant. For these two intentions, intention A can be the father intention, but for other intentions, intention A can be the child intention.

For one specific moment, we will start with an initial intention, and generate a sequence of intentions using the intention graph. The conversion between different intentions is bidirectional. Typically, the father intention generates the child intention, until we get the intentional action node and choose an action. However, sometimes, we need to recall. *E.g.*, when the child intention is finished, we will abandon this intention, and return to the father intention. Then, we will continue the recall, or generate a new child intention, until we get an intentional action node again.

Edge The edge between different intentions represents the condition when the conversion can be made. Theoretically, the condition should be estimated by the agent. However, here, for simplification, we will estimate the condition throughout the world.

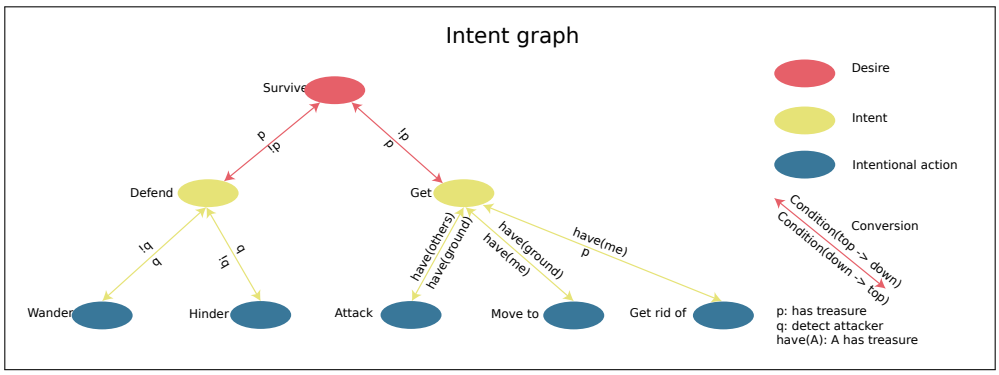

Figure 2: The intent graph we used. We have three different types of intention: pure intent(desire), intent, and intentional action. Pure intent is the start of the intent graph, which represents the tendency that our agents choose, like the desire for cooperation, or the final goal our agents want to achieve, like the desire for survival. The intent is the ordinary intention, and is not associated with the action. Intentional action is the intention associated with the action. Our agents will choose their actions that apply to the world based on intentional action.

3.2.2 Inference

The first question is where we should achieve inference. As mentioned above, all the agents have the same behavior model and intent graph. With comprehensive observation, the inference of others' intentions should be the same among all the agents. Thus, in practice, the world can achieve the inference process rather than the agents to reduce complexity.

The second question is how. The intent can be divided into pure intent(desire), intent, and intentional action, and each kind can have sources of different elements, which makes the general inference algorithm challenging to design. So we choose to design a specific algorithm that can solve the intent inference in our specific problem. Attention to object is what we use. Any Intentional action, if it can influence reality, should have entities to bear it. Thus, if we can infer the entities that the action involves, we can partly infer possible actions. As we only have move and hit action in the world, we only need to figure out which target the agent wants to move to, or move around, and which agent the agent wants to hit. The complete observation that our agents have makes the inference based on the perceptual attention of agents impossible. Instead, we choose to infer the intent based on moving attention. We try to get the target that the agent wants to move to from the direction the agent is moving. It should be a probability distribution function of the included angle between the line linking the agent and the target and the direction the agent is moving. Then we choose a threshold value. If the probability is larger than the value, we consider the entity as the target entity that the agent wants to move to.

After the process above, we got each agent's target entity. Then we can get the intentional action of each agent. The desire and intent can be inferred from the intentional action and the intent graph.

3.3 Action

The action is the available function for agents. There are two basic actions, moving and hitting. For moving, agents can move to their destination by changing their accelerated speed. This is fundamental for interaction in the environment. Hitting is based on moving, but can fetch down the treasure from the agent.

3.4 Ownership

Initialization Based on the work of Friedman [8], we used the principle of first possession to initialize the ownership. When generating the environment, we will compute the distance between every agent and every treasure. Then, the nearest agent to the treasure will own it.

Change The ownership can be changed based on the agents' actions. If there is only one agent around the treasure, then the ownership will be changed, and the agent near the treasure will own it.

Inference As we have stated above, all the agents have the same behavior model and comprehensive observation, which makes the inference the same among them. Thus in practice, the inference is achieved by the world rather than agents, to reduce the resource needed to compute. The world will contain the ownership relation at the beginning of the game, and update it if the agents' actions can change the relation. The agents will request the world for ownership information if needed.

4 Conclusion

In this paper, we modeled agents using the Theory of Mind (ToM) method, and studied the transition of intentions depending on the understanding of ownership. It established a unified computational framework for learning the virtual concepts, and we modeled the ownership concept of the agent based on it. What is more, the virtual concepts are not directly linked to actions, but to intents, making our model's generalization easier. Moreover, we use a hierarchical structure instead of atomic representation for modeling intents, which can reveal hidden relations between different intentions and provide a more straightforward method of inferring intention. We generated some data, conducted human experiments, and verified the model's validity and data based on this model. The experimental results indicate that this data set can better characterize, in part, people's learning and use of ownership concepts. Furthermore, we will improve our model in the future.

5 Discussion

5.1 Dark matters

Dark matters are the unobservable elements functioning in our daily lives. Instead of the observable elements, such as perception, attention, and action, dark matters play a more important role, such as intention, common ground, and causality. If we can not understand the dark matters of humans, we can hardly, if not never, build general artificial agents and make them cooperate with humans. Most dark matter has a biological base, which is a product of natural selection. The experiments conducted by developmental psychology support this argument[5]. They prove that infants understand most dark matters before they grasp language. It means that they all have a specific role in the evolution of humans, as natural selection only cares about the element that can help us live in the world. Thus if we want to study what dark matter is and which role it is actually playing, we must first find in which circumstance it is selected by evolution. Thus in this paper, we try to study the beginning of ownership based on the prevention of others' actions to our possession, which follows the argument of Veblen [19].

5.2 Nonverbal action

Tomasello [17] argued that nonverbal communication is the base of verbal communication, as infants learn to communicate in a nonverbal way before they grasp language. For ownership, it is the same. As Ellis [5] has argued, children learn the concept of ownership before they can use language. Thus the concept of ownership origins from nonverbal behavior and biological base rather than language. To study ownership thoroughly, we should begin with nonverbal behavior. So in our world, the agents

can only move and hit instead of speaking. If we want to study dark matters, we should pay more attention to nonverbal behavior than language, as most originate from nonverbal behaviors rather than language.

5.3 Intention and action

Most recent works link ownership with actions rather than intentions, and some researchers argue that the social norms are reflected in reality by the prevention of actions. They are right. However, we do not understand others by actions, but by intentions[4]. So we should link social norms with intentions rather than actions. The corresponding actions will change in different situations for one specific kind of ownership relation, but the corresponding intention should be the same. *E.g.*, if the agent does not want others to touch his cup, he can hinder others physically, or tell them not to touch the cup in words. However, his intention is the same: preventing others from touching his cup.

5.4 How to represent intention

In related work, we have introduced different methods of representing intention: the atomic method and the hierarchical method. We can have different intentions at the same time. When we are playing, we can say that we want to play, have fun, or bond with others. We argue that the hierarchical method can easily reveal hidden relationships among different intentions. In this paper, we use the intent graph to represent the relations. If we choose one desire, we can generate sub-intent from the desire, and then generate intentional action. Here we can have at least three different intentions at one time, and the relation between them can be found in the intent graph.

5.5 Mind modularity problem

The mind modularity problem has been one of the debates over the Theory of Mind for a long time. In 1983, Jerry Fodor published a book titled *The Modularity of Mind* to explain exactly what a module is. Fodor's theory has two parts[7]. The first one is positive. It says that input systems, such as perception systems, are modular. However, the second one is negative, which says that central systems, such as systems involved in practical reasoning, are not modular. Some post-Fodorian researchers proposed the massive modularity theory, which means the mind is modular through and through[13].

The debate will continue until we have an adequate biological method to detect the brain and have clear definitions for abstract concepts. If we can understand how the brains work more specifically, we will have the chance to show how brains work in our daily life. Moreover, up to now, we can not give a precise definition to some abstract concepts, so we can not divide the functionality of our brains into several identical parts. If one day we can achieve all the two aspects we mentioned, then we will have to build the connection between the abstract concept and our brains' activity. Then we can figure out the problem.

5.6 The basic limitation

Until now, existing AI has relied heavily on data. This is very much against the human learning model. Now many AI researchers have noticed this stark difference[16]. Methods based on deep learning and reinforcement learning, which are mainstream, rely on data, but one can always extract complex paradigms from fewer data. We may need a paradigm shift in learning[23].

Moreover, many models are efficient but need help understanding. We do not trust these models because we need to know whether the model will get the correct answer. Explainable AI(XAI) is trying to make the AI models can be interpreted, which is very important to gain humans' trust. DARPA proposes the concept of explainable artificial intelligence(XAI) for the first time[9]. According to [2], "eXplainable AI (XAI) proposes creating a suite of ML techniques that 1) produce more explainable models while maintaining a high level of learning performance, and 2) enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners". Nevertheless, up to now, there are some limitations in the XAI models. There is still a long way to go.

6 Contribution

Model The physical foundation of our model, the actions and the picture of agents and objects, were contributed by Zian Wang. Zhihao Cao designed the intent graph and the transition condition. Zhihao Cao designed the function classes of worlds, agents, and objects, and Zian Wang designed related functions to achieve the actions. Zhihao Cao achieved the inference of intention, initialization, and inference ownership.

Paper Zian Wang contributed to the "Mind modularity problem" and "The basic limitation" sections in the discussion, and translated some of Zhihao Cao's ideas into English in the introduction. Zian Wang contributes to the appendix part. The last of the paper, including the picture used to illustrate, is contributed by Zhihao Cao. Zhihao Cao does grammar check.

7 Appendix

7.1 Framework: the physical layer

On the physical layer, we should focus on how to make the agent have the ability to do what it wants and make them act like it is in a physical world. Due to the project's requirements, we cannot use any open-source projects, so we have to design our "physical layer" from scratch. We use the Pygame module to build our physical framework, which is the basis of our ToM method. In the following paragraph, we will introduce the design of our physical framework.

7.1.1 Class Design

Our project has two kinds of agents: one is called a processor, and the other is called an attacker. The processor's goal is to protect an item that belongs to him, and the attacker's goal is to rob the item. The attacker can hold the item when he reaches it. Furthermore, the processor can use his body to collide with the attacker to protect the item. If the attacker holds the item when colliding, the item will get knocked off, which is an expression of a procession. Although there are two kinds of agents, from the physical layer, we do not care about the "mind" of the agents but the action of the agents. From this perspective, the two kinds of agents are similar, so we use class inheritance to design the agents.

We have a hierarchical design for the function of the class. We can achieve some relatively complex functions by combining and utilizing simple functions. In the game, we use physical concepts to describe the agents' actions and states. We use a triple (position, velocity, accelerated velocity) to realize it. Furthermore, the velocity and accelerated velocity are both vectors, and we split vectors into norms and directions for easy calculation. In detail, the position is to describe the agent's state. The velocity can describe the agent's tendency to motion. Furthermore, accelerated velocity is used to describe the action of the agent.

7.1.2 Function Design

Motion The motion function is one of the most important functions used to perform the agent status update. It obeys the fundamental physical principle. That is, we use accelerated velocity to update the velocity and use velocity to update the position of the agents. In other words, we use the agent's action to influence its tendency of motion and use the tendency of motion to influence its state and to achieve something.

Resistance We use resistance to express the cost of action, which is very necessary. In the principle of maximum expected utility, an agent makes rational decisions based on beliefs and desires to maximize its expected utility.Zhu et al. [23] The cost will make the agent move more rationally.

Purposeful movement The agents must have the ability to move purposefully, meaning they should be able to move toward a particular position. We can control not the velocity but the accelerated velocity to realize this.

Collision The collision function is designed based on the perfectly elastic collision. To achieve a better performance, we also define the relative weight of the agent. In this circumstance, we can let the processor collide with the attacker, the attacker will be knocked away, and the item will be knocked off if the attacker holds it.

References

[1] Gertrude Elizabeth Margaret Anscombe. *Intention*. Harvard University Press, 2000. 2

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 6

[3] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007. 2

[4] Gergely Csibra and György Gergely. obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans. *Acta psychologica*, 124(1):60–78, 2007. 6

[5] Lee Ellis. On the rudiments of possessions and property. *Social Science Information*, 24(1): 113–143, 1985. 1, 2, 5

[6] Maria Isabel Aldinhas Ferreira, Joao Silva Sequeira, Mohammad Osman Tokhi, E Kadar, and Gurvinder Singh Virk. A world with robots. In *International Conference on Robot Ethics*. Springer, 2017. 1

[7] Jerry A Fodor. *The modularity of mind*. MIT press, 1983. 2, 6

[8] Ori Friedman. First possession: An assumption guiding inferences about who owns what. *Psychonomic Bulletin & Review*, 15(2):290–295, 2008. 2, 5

[9] David Gunning. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017. 6

[10] Steven Holtzen, Yibiao Zhao, Tao Gao, Joshua B Tenenbaum, and Song-Chun Zhu. Inferring human intent from video by sampling hierarchical plans. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1489–1496. IEEE, 2016. 2

[11] Jin Joo Lee, Fei Sha, and Cynthia Breazeal. A bayesian theory of mind approach to nonverbal communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 487–496. IEEE, 2019. 2

[12] Jon L Pierce, Tatiana Kostova, and Kurt T Dirks. The state of psychological ownership: Integrating and extending a century of research. *Review of general psychology*, 7(1):84–107, 2003. 2

[13] Richard Samuels. 2 massively modular minds: evolutionary. *Evolution and the human mind: Modularity, language and meta-cognition*, page 13, 2000. 2, 6

[14] Aimee E Stahl, Daniela Pareja, and Lisa Feigenson. Early understanding of ownership helps infants efficiently organize objects in memory. *Cognitive Development*, 65:101274, 2023. 1, 2

[15] Zhi-Xuan Tan, Jake Brawer, and Brian Scassellati. Thats mine! learning ownership relations and norms for robots. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8058–8065, 2019. 2

[16] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011. 6

[17] Michael Tomasello. *Origins of human communication*. MIT press, 2010. 2, 3, 5

[18] Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22, 2009. 2

[19] Thorstein Veblen. The beginnings of ownership. *American Journal of Sociology*, 4(3):352–365, 1898. 1, 2, 5

[20] Yuanfei Wang, Fangwei Zhong, Jing Xu, and Yizhou Wang. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. *arXiv preprint arXiv:2111.09189*, 2021. 2

[21] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6801–6809, 2018. 2

[22] Amanda L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998. 2

[23] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 2, 6, 7