Biologically Inspired Cognitive Architectures (2015) 11, 10–21

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bica

RESEARCH ARTICLE

Higher-order theory of mind in the Tacit Communication Game

Harmen de Weerd [a,*], Rineke Verbrugge [a], Bart Verheij [a,b]

[a] *Institute of Artificial Intelligence, Faculty of Mathematics and Natural Sciences, University of Groningen, The Netherlands*
[b] *CodeX, Stanford University, California, United States*

Received 7 November 2014; accepted 7 November 2014

KEYWORDS
Theory of mind;
Depth of reasoning;
Learning;
Cognitive hierarchy;
Communication;
Simulation

Abstract
To understand and predict the behavior of others, people regularly reason about others' beliefs, goals, and intentions. People can even use this theory of mind recursively, and form beliefs about the way others in turn reason about the beliefs, goals, and intentions of others. Although the evolutionary origins of this cognitively demanding ability are unknown, the Vygotskian intelligence hypothesis suggests that higher-order theory of mind allows individuals to cooperate more effectively. In this paper, we investigate this hypothesis through the Tacit Communication Game. In this game, two agents cooperate to set up novel communication in which a Sender agent communicates the goal to the Receiver agent. By simulating interactions between agents that differ in their theory of mind abilities, we determine to what extent higher orders of theory of mind help agents to set up communication. Our results show that first-order and second-order theory of mind can allow agents to set up communication more quickly, but also that the effectiveness of higher orders of theory of mind depends on the role of the agent. Additionally, we find that in some cases, agents cooperate more effectively if they reason at lower orders of theory of mind.

Introduction

While engaging in everyday activities, people regularly make use of *theory of mind* (Premack & Woodruff, 1978), and reason about what others know and believe. For example, people reason about the reasons that others might have to behave the way they do, and distinguish actions that are intentional from those that are accidental. People are also able to use this ability recursively, and use *second-order theory of mind* to reason about the way others reason about beliefs and goals. This allows them to understand sentences such as 'Alice *believes* that Bob does not *know* that Carol is throwing him a surprise party', and

* Corresponding author.
E-mail address: hdeweerd@ai.rug.nl (H. de Weerd).

http://dx.doi.org/10.1016/j.bica.2014.11.010
2212-683X/

make predictions about how this knowledge influences the behavior of Alice.

The human ability to make use of higher-order (i.e. at least second-order) theory of mind has been demonstrated both in tasks that require explicit reasoning using second-order beliefs (Apperly, 2011; Perner & Wimmer, 1985) as well as in strategic games (Hedden & Zhang, 2002; Meijering, Van Rijn, Taatgen, & Verbrugge, 2011). However, the use of higher-order theory of mind appears to be a uniquely human ability; whether any non-human species is able to make use of theory of mind of any kind is under debate (Clayton, Dally, & Emery, 2007; Penn & Povinelli, 2007; Tomasello, 2009; Van der Vaart, Verbrugge, & Hemelrijk, 2012; Martin & Santos, 2014). This suggests that there may be specific settings in which the ability to reason about the unobservable mental states of others is evolutionarily advantageous. Such settings would support the emergence of higher-order theory of mind, despite the high cognitive demands of such an ability.

According to the Vygotskian intelligence hypothesis (Vygotsky & Cole, 1978; Moll & Tomasello, 2007), cooperative social interactions play a crucial role in the development of human cognitive skills such as theory of mind. Tomasello, Carpenter, Call, Behne, and Moll (2005) propose that the uniquely human aspects of cognition that require higher-order theory of mind, such as constructing shared goals and joint intentions, developed because of the need for social cooperation. They state that higher-order theory of mind allows individuals to achieve levels of cooperation beyond those that are achieved by individuals that are unable to reason about the minds of others, and that are therefore unable to construct shared goals and joint intentions.[1]

Although the Vygotskian intelligence hypothesis suggests that theory of mind is necessary in some cooperative settings, simulation studies have shown that many forms of cooperation can evolve using simple mechanisms (Boyd, Gintis, Bowles, & Richerson, 2003; Nowak, 2006; Sigmund, 2010; Gärdenfors, 2012; Van der Post, De Weerd, Verbrugge, & Hemelrijk, 2013). Many animals are known to engage in cooperative interactions without relying on higher-order theory of mind (Wilkinson, 1984; Dugatkin, 1997; Crespi, 2001; Tomasello, 2009). Even highly organized and complex cooperative behavior such as cooperative hunting by lions, wolves, and chimpanzees can be described using fixed roles that rely on simple cues, without the need to construct joint intentions (Tomasello et al., 2005; Tomasello, 2009; Muro, Escobedo, Spector, & Coppinger, 2011). In fact, even when cooperation is risky, simple punishment strategies are enough to stabilize cooperation (Boyd & Richerson, 1992).

Since many forms of cooperation can be stabilized without the use of theory of mind, we instead focus on the process of establishing and coordinating cooperation between agents. We investigate a particular form of cooperation through the Tacit Communication Game (De Ruiter, Noordzij, Newman-Norlund, Hagoort, & Toni, 2007; Newman-Norlund et al., 2009; Blokpoel et al., 2012). In this game, a pair of players needs to set up communication that allows one player to inform the other player about the goal of the game. We make use of agent-based computational models to investigate how higher-order theory of mind can help agents to set up communication. Agent-based models have previously been used to explore the origins of communication (Steels, 2003; Scott-Phillips, Kirby, & Ritchie, 2009; De Bie, Scott-Phillips, Kirby, & Verheij, 2010; Steels, 2011), as well as to determine the effectiveness of higher order of theory of mind in competitive settings (De Weerd, Verbrugge, & Verheij, 2013) and mixed-motive settings (De Weerd, Verbrugge, & Verheij, 2014). By simulating interactions between computational agents, we can determine how the theory of mind abilities of these agents influences their performance.

The remainder of this paper is set up as follows. In Section 'Theory of mind in communication', we present one particular way in which theory of mind can be helpful in communication. Section 'Game setting' introduces the Tacit Communication Game. In Section 'Theory of mind in the Tacit Communication Game', we describe the agent model and show how the ability to reason at higher orders of theory of mind changes the way agents play the Tacit Communication Game. The details and the results of the simulated interactions between agents are found in Section 'Results'. Section 'Discussion' discusses our results, and compares these results to related work.

Theory of mind in communication

The Vygotskian intelligence hypothesis suggests that there are forms of cooperation that support the emergence of higher-order theory of mind. One way in which theory of mind may be beneficial for social cooperation is through communication. Bidirectional optimality theory (Blutner, 2000) suggests that in order to establish proper communication, the speaker has to take into account the perspective of the hearer, while the hearer has to take into account the perspective of the speaker. This may require a recursive mechanism of bidirectional optimization similar to second-order theory of mind (Flobbe, Verbrugge, Hendriks, & Krämer, 2008). In this system, a hearer decides on the correct interpretation of a sentence by making use of second-order theory of mind, by considering how the speaker would predict the hearer to interpret the sentence. De Hoop and Krämer (2006) argue that an inability to optimize bidirectionally is the reason for children to make mistakes in the interpretation of the following Dutch sentence *S*:

S: Er ging twee keer een meisje van de glijbaan af.
there went two time a girl of the slide down
''Twice a girl went down the slide.''

Sentence *S* allows for two different interpretations. The sentence can be interpreted as a single girl that went down the slide twice, or two separate girls that went down the slide once. Although Dutch children accept both interpretations, adults prefer the interpretation in which two separate girls went down the slide once.

According to De Hoop and Krämer (2006), bidirectional optimization accounts for adults' interpretation of the indefinite subject *een meisje* ('a girl') in sentence *S*. The preferred interpretation is that the indefinite subject *een*

[1] For the computational complexity of joint intentions, see Dziubiński, Verbrugge, and Dunin-Kęplicz (2007).