Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 28 (2014) 12–19

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Mindreading deception in dialog

Action Editor: Paul Bello

Alistair M.C. Isaac [a], Will Bridewell [b,*]

[a] Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA 19104, United States
[b] Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305, United States

Available online 23 July 2013

Abstract

This paper considers the problem of detecting deceptive agents in a conversational context. We argue that distinguishing between types of deception is required to generate successful action. This consideration motivates a novel taxonomy of deceptive and ignorant mental states, emphasizing the importance of an ulterior motive when classifying deceptive agents. After illustrating this taxonomy with a sequence of examples, we introduce a Framework for Identifying Deceptive Entities (FIDE) and demonstrate that FIDE has the representational power to distinguish between the members of our taxonomy. We conclude with some conjectures about how FIDE could be used for inference.
Published by Elsevier B.V.

Keywords: Mental state ascription; Ulterior motives; Lying; False belief

1. Introduction

Every productive interaction between humans depends critically on the process of mindreading: the attribution of mental states to other agents. If another person can successfully infer your beliefs and desires, then that person can interact with you strategically, anticipating and preparing for your actions to compete or coordinate. For artificial agents to successfully participate in complex social interactions, they also will need strategies for mindreading.

Ideal mindreading requires multimodal input, including utterances, eye movements, body gestures, galvanic skin response, and other features. However, humans routinely interact successfully in limited cue conditions. For example, in an internet chat session with a stranger, we lack access to facial expressions, body language, and past experiences with that person. Instead, we must attribute beliefs and desires to the stranger based solely on the words typed and our general knowledge of such situations. A limited cue scenario such as this offers a delineated testing ground for mindreading frameworks for artificial agents.

A considerable amount of work exists that attempts to infer mental state from dialog content alone (e.g., Cohen & Perrault, 1979; Carberry & Lambert, 1999), but the bulk of this literature assumes agents are cooperative and speak veridically. In contrast, this paper introduces a general framework for mindreading from dialog with the express goal of detecting deception. More specifically, we take as our primary focus the problem of distinguishing between different types of deceptive and sincere speech. Our motivation is the observation that the correct response to an utterance differs depending on its categorization (as a lie, a sincere statement of false belief, a veridical statement uttered with deceptive intent, etc.). One might confront a liar, but gently correct a sincere but ignorant speaker. In some contexts, one might even decide to participate in a falsehood if it is strategically efficacious.

To crystallize the problem, we concentrate on the special case in which an agent believes some proposition P and his interlocutor utters not(P). In this situation, mindreading is critical for determining how beliefs about the other agent should be revised and deciding what further actions to

* Corresponding author.
E-mail addresses: aisaac@sas.upenn.edu (A.M.C. Isaac), will.bridewell@nrl.navy.mil (W. Bridewell).

1389-0417/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.cogsys.2013.07.001

take. Does the speaker utter not(P) sincerely? Is he lying? If the former, the agent might act to educate the speaker by providing his evidence for P; if the latter, we claim that the agent must infer the speaker's ulterior motive. The need to correctly calculate response on the basis of a conflict between a speaker's utterance and the hearer's belief appears in many conversational scenarios including police interrogations, court testimony, physician–patient interactions, political debates, and even water-cooler talk.

The literature on deception emphasizes a number of fine-grained distinctions for characterizing different attitudes an agent might have toward an utterance. In addition to lying and distinct from the sincere categories of false belief and ignorance, there are other less well studied forms of deception. For example, *paltering* involves speaking truthfully with the intent to deceive (Schauer & Zeckhauser, 2009), as when a car dealer (veridically) emphasizes the quality of a car's wheels to distract the buyer from a problem with the engine. *Bullshitting* involves speaking without either knowing or caring about the truth value of one's utterance (Frankfurt, 2005). A less discussed, but also interesting example is *pandering*, when the speaker does not care about the truth value of their utterance, instead speaking it solely because they believe the listener desires to hear it. Whereas the philosophical literature on deception has focused primarily on the moral status of these classes of utterances, we focus on how to distinguish among them in a limited-cue, conversational situation.

Our strategy differs from previous work in that we (1) categorize a broad variety of deceptive states within a unified framework, (2) emphasize the importance of ulterior motives rather than "intent to deceive," and (3) propose a representation designed to be rich enough to support the detection of deception. The claim that deception and sincere discourse are qualitatively distinct is uncontroversial. We go a step further and assert that attempts to treat bullshitting, lying, pandering, and paltering as a single activity will lead to as much confusion as treating false belief, ignorance, and veridical speech as one and the same. The efforts of Sakama, Caminada, and Herzig (2010) to logically define and analyze lying, bullshit, and paltering—which they call "deception"—support our assertion. However, their logical analysis turns on acceptance of the condition that liars intend the hearer to believe their fallacious utterance, which has been hotly disputed (Mahon, 2008). We sidestep that debate by relying on the more general concept of an ulterior motive: intuitively, a goal with higher priority than those goals implied by the conversational context. The intent to deceive may follow from such a goal, but the ulterior motive itself ultimately determines action. These considerations motivate the Framework for Identifying Deceptive Entities (FIDE). We argue that FIDE includes features both necessary and sufficient for a mindreading system to detect deception.

We structure the rest of this paper around FIDE. In the next section, we develop a classification of deceptive states, illustrating our distinctions with multiple interpretations of a short dialog. We argue for the necessity of representing the goals and beliefs of other agents, including their potential ignorance and ulterior motives, so as to fully characterize these distinctions. Section 3 introduces FIDE and a formalism for representing mental states. Section 4 illustrates the framework's expressive power by applying it to examples from Section 2. Success at representing our examples demonstrates the sufficiency of FIDE for capturing the important distinctions. We then discuss strategies for inferring deceptive states within a system that implements this framework. Finally, we summarize our argument and anticipate future research.

2. An example: the water cooler

To see more clearly the necessity of mindreading and reasoning about ulterior motives for dialog, consider the following simple exchange:

Scene: *Bartleby & Bartleby, LLP*
(Jones and Pratt stand next to a gurgling water cooler.)

Jones: So, I hear Smith is going to be promoted to VP.

Pratt: That's what you get for kissing old man Bartleby's ass.

Jones' response to Pratt will depend critically on his own mental attitude toward Smith's promotion. For the sake of argument, assume that Jones believes Smith's promotion to be merit-based. How then should he interpret and respond to Pratt's assertion that it was the result of cronyism?

The most socially generous interpretation of Pratt's statement is as a straightforward instance of *false belief*. For example, Pratt may have observed but misinterpreted conversations between Smith and Bartleby, forming the sincerely held but inadequately justified and ultimately incorrect belief that the promotion was an act of cronyism. In this circumstance, Pratt's utterance may have no motive behind it other than the straightforward Gricean mandate to speak the truth. If Jones infers that motive, he might respond by offering Pratt evidence that Smith's promotion was merit-based, working with him to realize their shared goal of reaching the truth of the situation.

At the opposite end of the spectrum is the interpretation of Pratt's utterance as a full-blown instance of *lying*. For example, Pratt may know full well that Smith's promotion is merit-based. His assertion to the contrary must then be based on some *ulterior motive*. For instance, Pratt may have the goal of bringing Jones to believe that Smith's promotion was undeserved. This goal is ulterior in the sense that it contradicts the default Gricean assumption that the purpose of conversation is to convey truth about the state of the world. As we shall see, the presence of an ulterior motive most thoroughly characterizes deceptive speech, not the speaker's attitude toward the truth value