Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 28 (2014) 20–43

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Nuclear deterrence and the logic of deliberative mindreading ☆

Action Editor: Paul Bello

Selmer Bringsjord [a,b,c,d,*], Naveen Sundar Govindarajulu [a,b,c], Simon Ellis [b,c], Evan McCarty [b,c], John Licato [b,c]

[a] Department of Cognitive Science, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA
[b] Department of Computer Science, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA
[c] Rensselaer AI & Reasoning (RAIR) Lab, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA
[d] Lally School of Management & Technology, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA

Available online 16 September 2013

Abstract

Although the computational modeling of "mindreading" (e.g., believing that you believe that there's a deadly boa in the box, Smith mindreadingly predicts that you will refrain from removing the top) is well-established, this success has been achieved primarily in connection with scenarios that, relatively speaking, are both simple and common. Herein, we introduce a new computational-logic framework that allows formalization of mindreading of a rather more demanding sort: viz., **deliberative** multi-agent mindreading, applied to the realm of nuclear strategy. This form of mindreading, in this domain, is both complex and uncommon: it for example can quickly involve at least formulae reflecting *fifth*-order beliefs, and requires precise deductive reasoning over such iterated beliefs. In addition, the relevant models and simulations involve three, four, five agents, and sometimes many more. In the nuclear-strategy realm, for example, the better kind of modeling, simulation, and prediction (MSP) that our framework is intended to enable, should ultimately be capable of formalizing, at once, the arbitrarily iterated beliefs of at least every civilized nation on Earth. Based on our modeling, we present a set of desiderata that any modeling framework should satisy to be able to capture deliberative multi-agent mindreading in domains such as nuclear deterrence. Using our desiderata, we evaluate game theory, metagame theory, digital games, and tabletop games when used to model nuclear deterrence. Finally, we consider and rebut possible objections to our modeling paradigm and conclude with a map of ongoing and future work.

Keywords: Nuclear deterrence; Deliberative mindreading; Deontic cognitive event calculus

1. Introduction

The computational modeling of "mindreading" (e.g., believing that you believe that there's a deadly boa in the box, Smith mindreadingly predicts that you will refrain from removing the top) is well-established (e.g., see Arkoudas & Bringsjord, 2009; Bello, Bignoli, & Cassimatis, 2007). However, past success has been achieved in connection with scenarios that, relatively speaking, are both simple and common. Consider for instance the false-belief scenario, which drives both of (Arkoudas & Bringsjord, 2009; Bello et al., 2007); a scenario in which a young child, or the computational agent serving as a simulacrum thereof, must predict where some other agent will look in order to retrieve an object from *b*. A successful prediction requires that the child believe that the other agent believes that the object is located in *b*. Everyday life, from toddlerhood to (lucid) senescence, is filled with the need to make such predictions in such two-agent cases, on the strength of such second-order beliefs. You do not need a snake

☆ We owe an immeasurable debt to two anonymous referees and the erudite issue editor for brilliant and spirited comments on and objections to earlier drafts of the paper. In addition, Bringsjord and Sundar G. express their gratitude for financial support provided by the John Templeton Foundation and AFOSR.

* Corresponding author at: Department of Cognitive Science, Rensselaer Polytechnic Institute, 110-8th Street, Troy, NY 12180, USA.
E-mail addresses: selmer@rpi.edu (S. Bringsjord), govinn@rpi.edu (N.S. Govindarajulu), elliss5@rpi.edu (S. Ellis), mccare4@rpi.edu (E. McCarty), licatj@rpi.edu (J. Licato).

1389-0417/$ - see front matter © 2014 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.cogsys.2013.08.001

and a box or other contrivances to exemplify the logical relationships: If Jones is standing beside Smith while the latter is cooking a meal, and the former is considerate, Jones will not want to be located so as to block Smith's removal of now-grilled chorizo from one pan in order to add it to the sauce in another—and the courtesy of Jones inheres in his second-order belief about what Smith believes. In addition, both the number and average syntactic complexity of the formulas required to model such scenarios is relatively small.

Herein, we introduce a new computational-logic framework that allows formalization of mindreading of a rather more demanding sort: viz., **deliberative** multi-agent mindreading, applied to the realm of nuclear strategy. This form of mindreading, in this domain, is both complex and uncommon: it for example can quickly involve at least formulae reflecting *fifth*-order beliefs, and requires precise deductive reasoning over such iterated beliefs. In addition, the relevant models and simulations involve three, four, five agents, and sometimes many more. In the nuclear-strategy realm, for example, the better kind of modeling, simulation, and prediction (MSP) that our framework is intended to enable, should ultimately be capable of formalizing, at once, the arbitrarily iterated beliefs of at least every civilized nation on Earth.

Our plan for the present paper: In the next section (Section 2), we use a highly expressive intensional logic ($\mathcal{DCEC}^*$), embedded within a turnstyle rubric, to model, in four increasingly robust ways, snapshots taken of a four-agent, real-world interaction relating to nuclear deterrence. (The four agents are idealized representatives of the U.S., Israel, Iran, and Russia.) As we explain, this modeling is undertaken with the purpose of achieving simulations that enable predictions about the future, conditional on what actions are performed before at least the end of the future to be charted. In Section 3, we use and extend our modeling in Section 2 to prove that the U.S.'s applying severe economic sanctions, under certain reasonable suppositions, will not deter Iran from working toward massive first-strike capability against Israel. After taking stock of the eight chief advantages of (= desiderata derived from) our modeling approach (Section 4), we explain that both modern digital and tabletop games, and game and metagame theory, are inadequate as a basis for such modeling (Section 5). Next, we anticipate and rebut a series of objections to our new paradigm (Section 6). Finally, in a brief concluding section, we point toward our ongoing and future work.

2. The scenario and our model thereof

In this section, we first present our logicist framework, $\mathcal{DCEC}^*$, and then consider increasingly complex models of nuclear deterrence represented in this framework. The first two models are simple, in that their structure is fixed and the only possibility of variation is through adjustment of parameter values. Specifically, there is no provision for incorporating deliberative mind-reading in these two models. The third model builds upon the first two and uses $\mathcal{DCEC}^*$ to specify the model, and accordingly has enough expressive power to capture mindreading by the players involved. In addition to mindreading, the third model can also capture any arbitrary scenario that could be of relevance. For example, the first two models are agnostic on whether communication between the U.S. and Israel could be monitored by Russia for Iran. If we want to look at the effects of Russia monitoring such communication, we could, in principle, supply a statement of this fact and other relevant information in the form of a set of statements Γ_{ND} to a semi-automated system of our proof calculus, and ask the system questions ϕ that we might be interested in (where ϕ contains information about the relevant deterrence scenario). We argue that the proof calculus should be expressive enough to model deliberative mindreading. This entails that the formal calculus contain, at a minimum, syntax for expressing intensional operators like *knows*, *believes*, *ought*, and for expressing time, change, events, and actions.

It's particularly important to realize that in modeling deterrence, we are ultimately interested in answering the following question via simulation:

$$\Gamma_{ND} \vdash_{\mathcal{DCEC}^*} happens(action(\mathsf{iran}, attack(\mathsf{israel})), \mathsf{T})?$$

In the following sections (Sections 2.1 and 2.2), $\mathcal{DCEC}^*$ will be presented and the above question will be explained in more detail.

2.1. $\mathcal{DCEC}^*$

$\mathcal{DCEC}^*$ (deontic cognitive event calculus) is a *multi-sorted quantified modal logic*[1] that has a well-defined syntax and a proof calculus. The syntax of the language of $\mathcal{DCEC}^*$ and the rules of inference for its proof calculus are shown in Fig. 1. $\mathcal{DCEC}^*$ syntax includes a system of sorts S, a signature f, a grammar for terms t, and a grammar for sentences ϕ; these are shown on the left half of the figure. The proof calculus is based on natural deduction (Jaśkowski, 1934), and includes all the introduction and elimination rules for first-order logic, as well as rules for the modal operators; the rules are listed in the right half of the figure.

The formal semantics for $\mathcal{DCEC}^*$ is still under development; a semantic account of the wide array of cognitive and epistemic constructs found in the logic is no simple task—especially because of two self-imposed constraints: resisting fallback to the standard ammunition of possible-worlds semantics (which for reasons beyond the scope of the present paper we find manifestly implausible as a technique for formalizing the meaning of epistemic operators), and resisting the piggybacking of deontic operators on

[1] Manzano (1996) covers muli-sorted first-order logic (MSL). Details as to how a reduction of intensional logic to MSL so that automated theorem proving based in MSL can be harnessed is provided in (Arkoudas & Bringsjord, 2009).