Nuclear Deterrence and the Logic of Deliberative Mindreading[*]

Selmer Bringsjord • Naveen Sundar G. • Simon Ellis • Evan McCarty • John Licato
Department of Cognitive Science
Department of Computer Science
Rensselaer AI & Reasoning (RAIR) Lab
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
{selmer,govinn,elliss5,mccare4,licatj}@rpi.edu

June 23, 2013

Contents

[*] 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.

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 Bello, Bignoli & Cassimatis 2007, Arkoudas & Bringsjord 2009). 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 (Bello et al. 2007, Arkoudas & Bringsjord 2009); 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 don't need a snake 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 (§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 (§4), we explain that both modern digital and tabletop games, and game and metagame theory, are inadequate as a basis for such modeling

(§5). Next, we anticipate and rebut a series of objections to our new paradigm (§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:

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

In the following 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 Figure 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 pre-established logics not expressly created and refined for the purpose of doing justice to moral reasoning in the human realm.[2] The issue of a formal semantics is taken up in more depth in section 6.2.

[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).

[2] Such piggybacking is the main driver of (Horty 2012), in which deontic logic is understood via aligning it with default logic.

Figure 1: Deontic Cognitive Event Calculus

Syntax

$S ::=$ Object | Agent | Self $\sqsubset$ Agent | ActionType | Action $\sqsubseteq$ Event | Moment | Boolean | Fluent | Numeric

$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$

$\phi ::=$ p : Boolean $\mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \phi \rightarrow \psi \mid \phi \leftrightarrow \psi \mid \forall x : S.\ \phi \mid \exists x : S.\ \phi$ $f ::=$
$\mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi)$
$\mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,holds(f,t')) \mid \mathbf{I}(a,t,happens(action(a^*,\alpha),t'))$
$\mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t'))$

action : Agent × ActionType → Action
initially : Fluent → Boolean
holds : Fluent × Moment → Boolean
happens : Event × Moment → Boolean
clipped : Moment × Fluent × Moment → *Boolean*
initiates : Event × Fluent × Moment → Boolean
terminates : Event × Fluent × Moment → Boolean
prior : Moment × Moment → Boolean
interval : Moment × Boolean
$*$: Agent → Self
payoff : Agent × ActionType × Moment → Numeric

Rules of Inference

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))}\ [R_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))}\ [R_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \leq t_1 \ldots t \leq t_n}{\mathbf{K}(a_1,t_1,\ldots \mathbf{K}(a_n,t_n,\phi)\ldots)}\ [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [R_4]$$

$$\frac{t_1 \leq t_3, t_2 \leq t_3}{\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1 \rightarrow \phi_2) \rightarrow (\mathbf{K}(a,t_2,\phi_1) \rightarrow \mathbf{K}(a,t_3,\phi_2)))}\ [R_5]$$

$$\frac{t_1 \leq t_3, t_2 \leq t_3}{\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1 \rightarrow \phi_2) \rightarrow (\mathbf{B}(a,t_2,\phi_1) \rightarrow \mathbf{B}(a,t_3,\phi_2)))}\ [R_6]$$

$$\frac{t_1 \leq t_3, t_2 \leq t_3}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \rightarrow \phi_2) \rightarrow (\mathbf{C}(t_2,\phi_1) \rightarrow \mathbf{C}(t_3,\phi_2)))}\ [R_7]$$

$$\frac{}{\mathbf{C}(t,\forall x.\ \phi \rightarrow \phi[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)}\ [R_9]$$

$$\frac{}{\mathbf{C}(t,[\phi_1 \land \ldots \land \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi])}\ [R_{10}]$$

$$\frac{\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\phi \rightarrow \psi)}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \land \phi)}\ [R_{11b}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}] \qquad \frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

$$\frac{\begin{array}{c}\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\mathbf{O}(a^*,t,\phi,happens(action(a^*,\alpha),t')))\\ \mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t'))\end{array}}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,happens(action(a^*,\alpha),t')))}\ [R_{14}]$$

$$\frac{\phi \leftrightarrow \psi}{\mathbf{O}(a,t,\phi,\gamma) \leftrightarrow \mathbf{O}(a,t,\psi,\gamma)}\ [R_{15}]$$

Of course, we have informal interpretations for the different "parts of speech" in $\mathcal{DCEC}^*$: We denote that agent a knows ϕ at time t by $\mathbf{K}(a,t,\phi)$. The operators $\mathbf{B}$ and $\mathbf{P}$ have a similar informal interpretation for belief and perception, respectively. $\mathbf{D}(a,t,holds(f,t'))$ says that the agent a at time t desires that the fluent f holds at time t'. The formula $\mathbf{S}(a,b,t,\phi)$ captures declarative communication of ϕ from agent a to agent b at time t. Public declaration of ϕ at time t by a is denoted by $\mathbf{S}(a,t,\phi)$. Common-knowledge of ϕ in the system at time t is denoted by $\mathbf{C}(t,\phi)$. Common-knowledge of some proposition ϕ holds exactly when every agent knows ϕ, and every agent knows that every agent knows ϕ, and so on *ad infinitum*. Note the restrictions on the form of $\mathbf{I}$ and $\mathbf{O}$. $\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))$ indicates that the agent a at time t intends to perform an action of type α at some time t'; the $*$ operator, written here in postfix form, ensures that this is an exact self-referential attitude, and not an attitude that happens to hold of the same agent by happenstance. This representation closely follows that of Castañeda, and a more elaborate account of self-reference in $\mathcal{DCEC}^*$can be found in (Bringsjord & Govindarajulu 2013). The latest addition to the calculus is the *ought-to-be* dyadic deontic operator $\mathbf{O}$ presented in (Goble 2003, McNamara 2010), intended to help dodge Chisholm's paradox, which plagues Standard Deontic Logic (SDL).[3] $\mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t'))$ is to be read as: *"if it is the case that a at time t believes ϕ then that α is obligatory for a and this is known by a*." The Moment sort is used for representing time-points. We assume that time-points are isomorphic with $\mathbb{N}$; and function symbols (or functors) $+,-$; and relation symbols $>,<,\geq,\leq$ are available, under standard interpretations.

$\mathcal{DCEC}^*$has a classical monotonic view of the knowledge of the world possessed by agents, and hence regards knowledge to be unchanging. This means that if an agent knows ϕ at some time t, then the agent will continue to know ϕ for all time.[4] Beliefs possessed by an agent can change as time passes and events happen and fluents change, but knowledge remains constant or increases. This view of knowledge underpins all inference rules that have a knowledge component. For example, $[R_3]$ states that if some information ϕ is common knowledge at a certain time, then we can derive that an agent knows that at a certain time that another agent knows at another time, and so on, until finally, that an agent knows ϕ at a time, with all these time indexicals occurring later than the moment at which the common knowledge holds. Rule R_{15} for $\mathbf{O}$ is based on the only rule for the ought-to-be operator in (Goble 2003, McNamara 2010), which can be easily and plausibly interpreted as also holding for the ought-to-do case. Rule R_{14} connects the $\mathbf{O}$ operator with the knowledge and belief operators. The rule is to be informally read as follows: *"If it is the case that an agent believes that the agent ought to α when ϕ holds at any time, and it is the case that the agent ought to α when ϕ, and the agent believes that ϕ holds at a given time, then the agent knows that the agent intends to perform action α.".*[5]

[3] An excellent overview of SDL and Chisholm's Paradox available in the Stanford Encyclopedia of Philosophy; see http://plato.stanford.edu/entries/logic-deontic.

[4] A critic might be tempted to object with a scenario that seemingly leads to an inconsistency via our doctrine of knowledge immutability. For example:

> Suppose that an agent perceives a door being open at time t_1, and perceives the door being closed at t_3. On the framework you propose, modeling of this includes: $\{\mathbf{P}(a,t_1,open(door)),\mathbf{P}(a,t,\neg open(door))\}$. But this leads to a contradiction, as one can derive $open(door)\wedge\neg open(door)$.

On closer examination, the modeling error is immediately visible: When one asserts $\mathbf{P}(a,t_1,open(door))$, this is equivalent to asserting that the agent perceives something which is true for all time ($c=c$ is one such statement), i.e., a door which is open independent of time. The correct formulae in our framework would be $\{\mathbf{P}(a,t_1,holds(open(door),t_1)),\mathbf{P}(a,t_3,\neg holds(open(door),t_3))\}$, which is perfectly benign.

[5] The complexity of R14 is perhaps well beyond what is needed for present purposes (the mere introduction of our framework and general methodology). For example, it would be simpler, and probably defensible in the context of the present prolegomenon, to simply go with $\dfrac{\mathbf{B}(a,t,\phi)\quad\mathbf{O}(a,t,\phi,\gamma)}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,\gamma))}$, where $\gamma\equiv happens(action(a^*,\alpha),t')$.

The rule for the communication operator is from the analysis in (Wooldridge 2009, Chapter 7); the rules for the rest of the modal operators come from (Arkoudas & Bringsjord 2008*a*) and (Bringsjord & Govindarajulu 2013). Rules R_{11a} and R_{11b} enable an agent which believes in $\{\phi_2, \ldots, \phi_n\}$ to also believe ψ if $\{\phi_1, \ldots, \phi_n\} \vdash \psi$. Some readers may be uncomfortable with the duo of R_{11a} and R_{11b}, but we have included them as this is not only realistic but also necessary when agents represent nations.

$\mathcal{DCEC}^*$ includes the signature of the classic Event Calculus ($\mathcal{EC}$) (Mueller 2006),[6] and the axioms of EC are assumed to be common knowledge in the system. EC is a first-order calculus that lets one reason about events that occur in time and their effects on fluents. $\mathcal{DCEC}^*$ includes, in addition to symbols for basic arithmetic functions and relations $S_{ar} = \{0, 1, +, ., <, >\}$, a relevant theory of arithmetic Φ_{arith}. The details vary from application to application; the power of this theory can be altered. This allows us to model agents that can access numerical propositions and calculations; for example, we might model an agent that uses simple utilitarian calculation to ground its decisions. The agents are also assumed to have some basic knowledge of causality going beyond the EC; this is expressed as common knowledge. These axioms are not pertinent to the present study; they can be found in (Arkoudas & Bringsjord 2008*b*).

$\mathcal{DCEC}^*$ has a set of distinguished constant symbols corresponding to when and by whom the reasoning is carried out: now is a symbol indicating the current time, and I is a symbol indicating the agent carrying out the reasoning. These features have not yet been fully developed, but the need for these features is explained in (Bringsjord & Govindarajulu 2013). These features are crucial in any formal system that seeks to be used by nation states formally modeling what they ought to do from a first-person perspective dynamically, taking account of how things work informally in the "real world." For example, if the U.S. wants to model what it should do in the future given information about things that have happened in the past, the U.S. can produce a series of models in the form of formal statements making use of I (or possibly a new symbol we) and now. Such modeling could be done in the third-person discarding the I symbol, but it is hard to see how any formal model could strive to be mature and user-friendly without properly accounting for when the model itself was formulated. For example, we show in our sample formal model in Section 2.2.3, that an informal premise $\mathcal{A}_6$ about what Iran knows about the future is easily cast into formal form as $\mathcal{A}_6$ using the now symbol.

2.2 Dynamic Models of Capability Development and Deterrence

We turn now to the promised scenario: a small snapshot of the overall interaction we are interested in ultimately modeling and predicting. The scenario involves four countries intertwined in a high-stakes nuclear drama unfolding before our eyes today: the United States, Iran, Israel, and Russia. We denote members of the quartet by: us, iran, israel, and russia.

Existing models of nuclear deterrence, such as those based solely on the concept of mutually assured destruction, are static, and assume that actors already have fully developed nuclear capabilities. Real-life scenarios are rather more complex, as shown by the flurry of statements released by Israel and the U.S. — statements which, taken at face value anyway, unambiguously convey a desire on the part of both nations that Iran not acquire first-strike capability relative to Israel (and, needless to say, relative to the mainland U.S.). Formal frameworks for multi-agent modeling of nuclear strategy should allow that modeling to capture: the

[6] Let the infix $t_1 < t2$ stand for $prior(t_1, t_2)$; then the axioms of the EC are:

$$\forall f : \text{Fluent},\ t : \text{Moment}.\ initially(f) \land \neg clipped(0, f, t) \Rightarrow holds(f, t)$$

$$\forall e : \text{Event},\ f : \text{Fluent},\ t_1, t_2 : \text{Moment}.\ happens(e, t_1) \land initiates(e, f, t_1) \land t_1 < t_2 \land \neg clipped(t_1, f, t_2) \Rightarrow holds(f, t_2)$$

$$\forall t_1, t_2 : \text{Moment}, f : \text{Fluent}.\ clipped(t_1, f, t_2) \Leftrightarrow (\exists e : \text{Event},\ t : \text{Moment}.\ happens(e, t) \land t_1 < t < t_2 \land terminates(e, f, t))$$

"temporally extended" effort on the part of nations to actively enhance their nuclear capability in sequential "ascending" steps well short of all-out first-strike capability; the masking of their intentions; other nations trying to prevent them from acquiring first-strike capability; and so on. Towards this end, we now present a series of dynamic models of nuclear strategy.

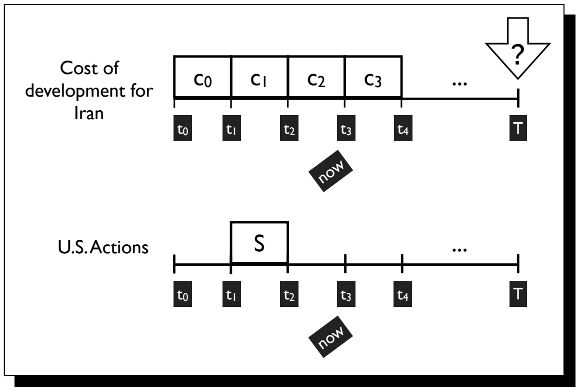
The general structure of the model represents nations by agents which seek to maximize payoffs resulting from actions. Agents perform actions at discrete time intervals. The time structure, recall, is assumed to be isomorphic to the set of natural numbers. The interaction or game starts at time 0, proceeds to $1, 2, \ldots$, and ends at some time T. There is at least one aggressor state, Iran, which seeks to develop nuclear capability and use it against another state, Israel. (In future models, other aggressors could easily be added.) There is a "coercer" state, the U.S., that actively seeks to prevent the aggressor from using its nuclear capability. The coercer has in its arsenal a set of graded options, ranging from economic sanctions to tactical strikes and outright decimation of the aggressor's capabilities. Each agent then seeks to maximize its payoff at time T. In this paper we look at the aggressor's payoff and consider detailed modeling of the aggressor.

Again, we start with two simple models (no mindreading) and end with a sketch of a preliminary model in $\mathcal{DCEC}^*$ that incorporates elements of deliberative mindreading.

2.2.1 Model 1: Fixed Costs of Development

Model 1 is simple, but dynamic in nature. We assume that the cost of development of nuclear capability for Iran is fixed per unit time at c, and that the total cost of the U.S.'s actions aimed at Iran is fixed at S. Let the benefit to Iran of attacking Israel with a full-blown nuclear strike be Δ. The U.S. could either intervene or not, and in response to U.S. action, Iran could either attack Israel by developing its capabilities up to time T, or not attack and stop developing its capabilities at time t'. Figure 2 illustrates the time structure of this model. Table 1 gives the payoff matrix for Iran.

Figure 2: Model 1



		Coercer Acts?	
		Yes	No
Aggressor Attacks?	Yes	$\Delta - S - c\mathsf{T}$	$\Delta - c\mathsf{T}$
	No	$-S - ct'$	$-ct'$

Table 1: Model 1 of Dynamic Deterrence

It can be seen that to deter Iran from attacking it must be that $c > \Delta/(\mathsf{T}-t')$. That is, the cost of developing more capability should be above a threshold. Another assumption is that if Iran stops at time t', it is in response to U.S. actions at or before time t'. As $t' \to \mathsf{T}$, the threshold $\Delta/(\mathsf{T}-t') \to \infty$. A non-obvious conclusion thus emerges: The more delayed the actions of the U.S., the more attractive it is for Iran to attack, even if the intensity of the actions by the U.S. are the same. By simply performing its interventions earlier in time, the U.S. can magnify the effects of its actions. In short, the U.S. should attack before

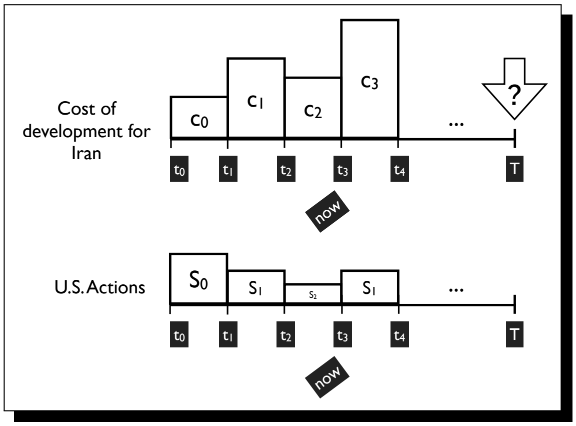
$$t' < \mathsf{T} - \Delta/c.$$

We see the following in Model 1: The more time T it takes for Iran to develop its capability, the more the U.S. can delay; the more Iran wants to attack Israel, the less the U.S. should delay its actions; and the cheaper it is for Iran to enhance its capability, the sooner the U.S. should act.

2.2.2 Model 2: Changing Costs

The previous model is obviously unrealistic; it serves here as merely a starting point to help us grasp the basic elements in play. In the next model, the cost of development for Iran changes continuously. This could be due to natural economic factors or actions by the U.S., overt or covert. The cost at time-point i is c_i. The actions taken by the U.S. are somewhat more "fine-grained" and vary across time: action taken by the U.S. at time i has cost S_i for Iran. The payoff matrix now changes to the following:

Figure 3: Model 2



		Coercer Acts?	
		Yes	No
Aggressor Attacks?	Yes	$\Delta - \sum_{i=0}^{\mathsf{T}} S_i - \sum_{i=0}^{\mathsf{T}} c_i$	$\Delta - \sum_{i=0}^{\mathsf{T}} c_i$
	No	$-\sum_{i=0}^{t'} S_i - \sum_{i=0}^{t'} c_i$	$-\sum_{i=0}^{t'} c_i$

Table 2: Model 2 of Dynamic Deterrence

Each c_i is dependent on c_{i-1} and S_{i-1}. Given this model, Iran will be deterred if

$$\sum_{i=t'+1}^{\mathsf{T}} c_i > \Delta - \sum_{i=t'+1}^{\mathsf{T}} S_i.$$

Assuming that the U.S. plans to respond with actions that have fixed total cost over a certain time period, the earlier these actions are performed the more effective they will be.

These two models leave a lot to be desired. For instance, it is not clear that if Iran stops developing its capability at time t' without any major action from the U.S. (e.g., as in the last column in Table 1), this behavior is in response to perceived threats from the U.S. The agents lacking full information about future actions of other agents have to resort to deliberative mindreading. There is no formal machinery in these two models to represent any mindreading that agents have to resort to. The next model shows how the apparatus of the $\mathcal{DCEC}^*$ can be used to build a theory of deliberative mindreading for nuclear deterrence.

2.2.3 Model 3: Cognitive Elements

The two models developed so far have assumed that the agents are perfect mechanical information-processors, in that they adhere to the model perfectly without any variation, and have access to all the relevant information, which makes mindreading unnecessary. The models can be made more realistic by introducing cognitive agents who do not have access to perfect and complete information about what the other agents are thinking; this makes mindreading essential. These agents are agents in $\mathcal{DCEC}^*$. They perform actions according to the dynamic model developed above, but the costs involved and the actions of other agents are not known in advance: the agents have beliefs about the costs. Agents can manipulate these beliefs by public communication of threats and other actions. In short, mindreading becomes essential, and makes the model move toward what is real.

The relevant $\mathcal{DCEC}^*$ agent symbols corresponding to the states are: $\{\mathsf{us}, \mathsf{iran}, \mathsf{russia}, \mathsf{israel}\}$. The U.S. has ability to perform an action of the following type: $\mathit{deter} : \mathsf{Numeric} \rightarrow \mathsf{ActionType}$, which means an action of a specific intensity to deter Iran performed by the U.S. If $\bar{c}$ is a $\mathcal{DCEC}^*$ term that denotes the natural number c, we can take $\mathit{deter}(\bar{0})$ to be a moderate economic action, $\mathit{deter}(\bar{1})$ to be a severe economic sanction, $\mathit{deter}(\bar{2})$ to mean limited military action, and so on. For convenience, we use the sloppier $\mathit{deter}(c)$ hereafter. Iran can enhance its capability at a certain cost $\mathit{enhance} : \mathsf{ActionType}$. Iran can also attack Israel: $\mathit{attack} : \mathsf{Agent} \rightarrow \mathsf{ActionType}$. The cost of enhancing Iran's nuclear capability at any time is given by $\mathit{cost} : \mathsf{Moment} \rightarrow \mathsf{Numeric}$. In addition to the above explicit actions, the participants can engage in communication at any time t of declarative information ϕ between themselves, modeled by the $\mathbf{S}(a,b,t,\phi)$ operator; or public communication, modeled by $\mathbf{S}(a,t,\phi)$.[7] The fluent *capable* denotes whether Iran's capacity is capable of being morphed into attack capability against Israel by time T. The fluent *destroyed* denotes whether Israel has been destroyed.

Let the set of $\mathcal{DCEC}^*$ formulae capturing our *nuclear deterrence* model be denoted by Γ_{ND}. Then, again, the ultimate prediction that we are interested in is of this form:

$$\boxed{\Gamma_{ND} \vdash \mathit{happens}(\mathit{action}(\mathsf{iran}, \mathit{attack}(\mathsf{israel})), \mathsf{T})?}$$

[7] Note that the two forms of **S** are in fact a different kind of syntactic object, but are represented by a similar letter for convenience. We have not included a group-communication operator and we acknowledge that we are ignoring intricate subtleties involved in communication.

Now we turn our attention to Γ_{ND}. What are its contents? A simple but plausible model that can be used to populate Γ_{ND} is given below in English.

$\boxed{A_0}$ If the U.S. does not desire that Israel be destroyed, and if the U.S. believes that Iran's attacking Israel will destroy Israel, then the U.S. ought to engage in actions that it believes will prevent this.

$\boxed{A_1}$ If the U.S. does not desire that Israel be destroyed, and if the U.S. believes that Iran's attacking Israel will destroy Israel, then the U.S. believes that it ought to engage in actions that it believes will prevent this.

$\boxed{A_2}$ At any time t, Iran intends to attack Israel at time T only if it believes that the benefits it gains by attacking, minus the cost of U.S. action from t to T, are greater than the cost of development from time t to T.

$\boxed{A_3}$ The U.S. knows A_2 above.

$\boxed{A_4}$ Iran knows A_3.

$\boxed{A_5}$ The cost of development at time t is $cost(t)$, and $cost(t_1,t_2)$ is the cost of development from time t_1 to time t_2. The sum of the costs to Iran of deterrence actions by the U.S. from time t_1 to time t_2 is denoted by $deter_{sum}(t_1,t_2)$.

$\boxed{A_6}$ At the current time, Iran only believes and does not know what future actions performed by the U.S. will be, and what future costs of development will be.

$\boxed{A_7}$ The U.S. publicly communicates what deterrence action it will perform in the immediate future.

$\boxed{A_8}$ The U.S. privately communicates to Israel what deterrence action it will perform in the immediate future.

$\boxed{A_9}$ Russia can monitor private communication between the U.S. and Israel regarding what the U.S. intends in the immediate future for deterring Iran. Russia then relays this information to Iran.

$\boxed{A_{10}}$ Russia believes that the U.S. does not know that Russia monitors its private communication with Israel.

$\boxed{A_{11}}$ Iran believes whatever information Russia supplies to it about the intentions of the U.S. communicated to Israel with regard to Iran.

$\boxed{A_{12}}$ The U.S. knows A_{11}.

$\boxed{A_{13}}$ If at any time Iran believes that it will be counterproductive to attack Israel, it stops enhancing its capability.

$\boxed{A_{14}}$ Iran does not attack Israel if it did not enhance at any particular time.

$\boxed{A_{15}}$ Iran believes that U.S. is capable of realizing its intentions.

$$\vdots$$

The timepoints in this model can correspond to different levels of granularity that one might be interested in. For short-term nuclear strategy, the time-points could be months; for long-term strategy, they could be years. So, though the above model is brutally incomplete, at least time-wise there is much flexibility for the future — and of course the flexibility of our framework extends in many other directions. At this point we proceed to the English assumptions $\boxed{A_i}$ translated into their formal counterparts $\mathcal{A}_i$, immediately below.
Note:

$$cost_{sum}(a,b) = \sum_{i=a}^{b} c_i \text{ and } deter_{sum}(a,b) = \sum_{i=a}^{b} S_i,\ c_i \text{ and } S_i$$

are such that

$$cost(t_i) = c_i \text{ and } happens(action(\mathsf{us}, deter(S_i)), t_i)$$

hold.

$\boxed{\mathcal{A}_0}$

$$\begin{array}{l}
\forall t,t' : \mathsf{Moment}\ \neg\mathbf{D}(\mathsf{us},t,\mathit{holds}(\mathit{destroyed}(\mathsf{israel}),t')) \Rightarrow\\
\mathbf{O}\left(\mathsf{us},t,\left(\begin{array}{l}
\mathit{initiates}(\mathit{action}(\mathsf{iran},\mathit{attack}(\mathsf{israel})),\mathit{destroyed}(\mathsf{israel}),\mathsf{T})\wedge\\
\mathit{terminates}(\mathit{action}(\mathit{us}^*,\mathit{deter}(v)),\mathit{capable},t')\wedge\\
\neg\mathit{holds}(\mathit{capable},t') \iff \neg\mathit{happens}(\mathit{action}(\mathsf{iran},\mathit{attack}(\mathsf{israel})),\mathsf{T})
\end{array}\right)\right.\\
\qquad\left.\mathit{happens}(\mathit{action}(\mathsf{us}^*,\mathit{deter}(v)),t')\right)
\end{array}$$

$\boxed{\mathcal{A}_1}$

$$\begin{array}{l}
\forall t,t' : \mathsf{Moment}\ \neg\mathbf{D}(\mathsf{us},t,\mathit{holds}(\mathit{destroyed}(\mathsf{israel}),t')) \Rightarrow\\
\mathbf{B}\left(\mathsf{us},t,\mathbf{O}\left(\mathsf{us},t,\left(\begin{array}{l}
\mathit{initiates}(\mathit{action}(\mathsf{iran},\mathit{attack}(\mathsf{israel})),\mathit{destroyed}(\mathsf{israel}),\mathsf{T})\wedge\\
\mathit{terminates}(\mathit{action}(\mathit{us}^*,\mathit{deter}(v)),\mathit{capable},t')\wedge\\
\neg\mathit{holds}(\mathit{capable},t') \iff \neg\mathit{happens}(\mathit{action}(\mathsf{iran},\mathit{attack}(\mathsf{israel})),\mathsf{T})
\end{array}\right)\right.\right.\\
\qquad\left.\left.\mathit{happens}(\mathit{action}(\mathsf{us}^*,\mathit{deter}(v)),t')\right)\right)
\end{array}$$

$\boxed{\mathcal{A}_2}$

$$\forall t : \mathsf{Moment}\left(\begin{array}{l}
\mathbf{I}(t,\mathsf{iran},\mathit{happens}(\mathit{action}(\mathsf{iran}^*,\mathit{attack}(\mathsf{israel})),\mathsf{T}))\\
\iff\\
\mathbf{B}(\mathsf{iran},t,\mathit{cost}_{\mathit{sum}}(t+1,\mathsf{T}) < \Delta - \mathit{deter}_{\mathit{sum}}(t+1,\mathsf{T}))\\
\hfill \wedge t < \mathsf{T}
\end{array}\right)$$

$\boxed{\mathcal{A}_3}$

$$\forall t' : \mathsf{Moment}\ \mathbf{K}\left(\mathsf{us},t',\left(\forall t : \mathsf{Moment}\left(\begin{array}{l}
\mathbf{I}(t,\mathsf{iran},\mathit{happens}(\mathit{action}(\mathsf{iran}^*,\mathit{attack}(\mathsf{israel})),\mathsf{T}))\\
\iff\\
\mathbf{B}(\mathsf{iran},t,\mathit{cost}_{\mathit{sum}}(t+1,\mathsf{T}) < \Delta - \mathit{deter}_{\mathit{sum}}(t+1,\mathsf{T}))\\
\hfill \wedge t < \mathsf{T}
\end{array}\right)\right)\right)$$

$\boxed{\mathcal{A}_4}$

$$\begin{array}{l}
\forall t'' : \mathsf{Moment}\\
\mathbf{K}\left(\mathsf{iran},t'',\left(\forall t' : \mathsf{Moment}\ \mathbf{K}\left(\mathsf{us},t',\left(\forall t : \mathsf{Moment}\left(\begin{array}{l}
\mathbf{I}(t,\mathsf{iran},\mathit{happens}(\mathit{action}(\mathsf{iran}^*,\mathit{attack}(\mathsf{israel})),\mathsf{T}))\\
\iff\\
\mathbf{B}(\mathsf{iran},t,\mathit{cost}_{\mathit{sum}}(t+1,\mathsf{T}) < \Delta - \mathit{deter}_{\mathit{sum}}(t+1,\mathsf{T}))\\
\hfill \wedge t < \mathsf{T}
\end{array}\right)\right)\right)\right)\right)
\end{array}$$

$\boxed{\mathcal{A}_5}$

$$\begin{aligned}
\mathit{cost}(t_1,t_2) &= \mathit{cost}(t_2) + \mathit{cost}(t_1,t_2-1)\\
\mathit{cost}(t,t) &= \mathit{cost}(t)\\
\mathit{deter}(t_1,t_2) &= S + \mathit{deter}(t_1,t_2-1) \wedge \mathit{happens}(\mathit{action}(\mathit{us},\mathit{deter}(S)),t_1)
\end{aligned}$$

$\boxed{\mathcal{A}_6}$

$$\begin{array}{l}
\forall t_f : \mathsf{Moment}\ t_f > \mathsf{now} \rightarrow\\
\left(\begin{array}{l}
\exists \mathsf{estimate} : \mathsf{Numeric}\ \mathbf{B}(\mathsf{iran},\mathsf{now},\mathit{happens}(\mathit{action}(\mathsf{us},\mathit{deter}(\mathsf{estimate})),t_f))\wedge\\
\forall \mathsf{estimate} : \mathsf{Numeric}\ \neg\mathbf{K}(\mathsf{iran},\mathsf{now},\mathit{happens}(\mathit{action}(\mathsf{us},\mathit{deter}(\mathsf{estimate})),t_f))
\end{array}\right)
\end{array}$$

$\boxed{\mathcal{A}_7}$

$$\forall t : \text{Moment}\ \exists v : \text{Numeric}\ \mathbf{S}(\text{us}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))$$

$\boxed{\mathcal{A}_8}$

$$\forall t : \text{Moment}\ \exists v : \text{Numeric}\ \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, happens(action(\text{us},^* deter(v)), t+1)))$$

$\boxed{\mathcal{A}_9}$

$$\begin{array}{c}\forall t : \text{Moment}\ \exists v : \text{Numeric}\ \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))) \rightarrow \\ \left(\begin{array}{c} \mathbf{P}(\text{russia}, t, \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1))))\wedge \\ \mathbf{S}(\text{russia}, \text{iran}, t, \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))) \end{array}\right)\end{array}$$

$\boxed{\mathcal{A}_{10}}$

$$\begin{array}{l}\forall t'' : \text{Moment} \\ \mathbf{B}\left(\text{russia}, t'', \neg\forall t' : \text{Moment}\ \mathbf{K}\left(\text{us}, t', \left(\begin{array}{l} \forall t : \text{Moment}\ \exists v : \text{Numeric} \\ \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1))) \\ \rightarrow \\ \mathbf{P}(\text{russia}, t, \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))) \end{array}\right)\right)\right)\end{array}$$

$\boxed{\mathcal{A}_{11}}$

$$\begin{array}{l}\forall t : \text{Moment}\ \forall v : \text{Numeric}\ \mathbf{S}(\text{russia}, \text{iran}, t, \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))) \\ \rightarrow \mathbf{B}(\text{iran}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))\end{array}$$

$\boxed{\mathcal{A}_{12}}$

$$\begin{array}{l}\forall t' : \text{Moment} \\ \mathbf{K}\left(\text{us}, t', \left(\begin{array}{l}\forall t : \text{Moment}\ \forall v : \text{Numeric}\ \mathbf{S}(\text{russia}, \text{iran}, t, \mathbf{S}(\text{us}, \text{israel}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))) \\ \rightarrow \mathbf{B}(\text{iran}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1)))\end{array}\right)\right)\end{array}$$

$\boxed{\mathcal{A}_{13}}$

$$\forall t : \text{Moment} \left(\begin{array}{l}\mathbf{B}(\text{iran}, t, cost_{sum}(t+1, \mathsf{T}) > \Delta - deter_{sum}(t+1, \mathsf{T})) \wedge t < \mathsf{T} \iff \\ \neg happens(action(\text{iran}, enhance), t)\end{array}\right)$$

$\boxed{\mathcal{A}_{14}}$

$$\begin{array}{c}\neg happens(action(\text{iran}, attack(\text{israel})), \mathsf{T}) \iff \\ \exists t : \text{Moment}\ t < \mathsf{T} \wedge \neg happens(action(\text{iran}, enhance), t)\end{array}$$

$\boxed{\mathcal{A}_{15}}$

$$\begin{array}{c}\forall t : \text{Moment}\ \forall v : \text{Numeric}\ \mathbf{B}(\text{iran}, t, \mathbf{I}(\text{us}, t, happens(action(\text{us}^*, deter(v)), t+1))) \iff \\ \mathbf{B}(\text{iran}, t, happens(action(\text{us}, deter(v)), t+1))\end{array}$$

3 A Prediction of Deterrence Failure with Supporting Proof Sketch

We now consider a micro-scenario replete with mindreading in which deterrence fails. For the micro-scenario, let's assume the following time structure: There are five distinct time-points $\langle t_0, t_1, t_2, t_3, t_4\rangle$ with the obvious temporal ordering and with $\mathsf{T} = t_4$. We also stipulate that these are the only time-points in the game such that:

$$\forall t : \text{Moment}\ t < T \Rightarrow t = t_0 \vee t = t_1 \vee t_2 \vee t = t_3$$

One scenario under which deterrence can fail, no matter what the U.S. does, is when Iran has irrational beliefs. This can be modeled by having Iran value attacking Israel very highly; given that scenario, the following holds:[8]

$$\begin{gathered}
\mathit{payoff}(\mathsf{iran}, \mathit{attack}(\mathsf{israel}), \mathsf{T}) = \Delta \Rightarrow \\
\forall t : \mathsf{Moment}\ \forall v : \mathsf{Numeric}\ \mathbf{B}(\mathsf{iran}, t, \Delta > v)
\end{gathered}$$

The above formula states that Iran values attacking Israel above anything else. Obviously one could reasonably doubt that this is the case. A more refined model of Iran's beliefs includes the following formula, which says that Iran places a finite value on attacking Israel, this premium being somewhat more than the amount of effort it has put into developing capability sufficient to first-strike attack Israel.

$$\forall t : \mathsf{Moment}\ \exists v : \mathsf{Numeric}\ \mathbf{B}(\mathsf{iran}, t, \Delta > v + \mathit{cost}_{sum}(t_0, \mathsf{T})) \wedge v \neq 0$$

Based on the above belief and beliefs regarding what the U.S. will do in the future and what the U.S. has done in the past, Iran can come to believe the following proposition or its negation.

$$\mathit{cost}_{sum}(t+1, \mathsf{T}) < \Delta - \mathit{deter}_{sum}(t+1, \mathsf{T}))$$

One way that Iran could end up with the above belief is by monitoring what the U.S. says to Israel about future actions and further deliberative mindreading based on that. We will now show a micro-scenario in which deterrence fails based on mindreading by Iran and the U.S.

Before we look at the U.S., its beliefs, and what it ought to do, we can safely presume that the following statement about the effects of Iran attacking Israel is common knowledge in the system:

$$\forall t, t' : \mathsf{Moment}\ \mathbf{C}\left(t, \left(\begin{array}{r} \mathit{initiates}(\mathit{action}(\mathsf{iran}, \mathit{attack}(\mathsf{israel})), \mathit{destroyed}(\mathsf{israel}), t') \wedge \\ \neg \mathit{holds}(\mathit{capable}, t') \iff \neg \mathit{happens}(\mathit{action}(\mathsf{iran}, \mathit{attack}(\mathsf{israel})), \mathsf{T}) \end{array}\right)\right)$$

We can also assume that the U.S believes that economic deterrence actions alone will damage Iran's capability:

$$\forall t, t' : \mathsf{Moment}\ \mathbf{B}\left(\mathsf{us}, t, \left(\mathit{terminates}(\mathit{action}(us^*, \mathit{deter}(1)), \mathit{capable}, t')\right)\right)$$

From public statements released by the U.S., it can be asserted that the U.S. does not wish to see Israel destroyed:

$$\forall t, t' : \mathsf{Moment}\ \neg\mathbf{D}(\mathsf{us}, t, \mathit{holds}(\mathit{destroyed}(\mathsf{israel}), t'))$$

[8] We have not elaborated on the sub-theory needed in $\mathcal{DCEC}^*$ about agents, actions, payoffs, costs, etc. We have used different but simple such theories in the past. The details of this do not matter for the sake of mindreading during nuclear deterrence; we thus skip them in the interest of efficiency. One such detail is that every action has a payoff for an agent.

Using the above three statements and some intricate but conceptually straightforward reasoning using $\mathcal{DCEC}^*$ rules (including the rule R_{14} for the $\mathbf{O}$ operator) we can derive the following statement:

$$\forall t, t' : \mathsf{Moment}\ \mathbf{I}\left(\mathsf{us}, t, \big(happens(action(us^*, deter(1)), capable, t')\big)\right)$$

At each time-point $< T$, the U.S. communicates to Israel its intention of imposing economic sanctions on Iran. This follows from instantiations of $\boxed{\mathcal{A}_8}$. U.S. believes that a steady barrage of economic sanctions should force Iran away from attacking Israel. Deterrence using only economic sanctions is represented by $deter(1)$. We have the following formulae:

$$\mathbf{S}(\mathsf{us}, \mathsf{israel}, t_0, \mathbf{I}(\mathsf{us}, t_0(happens(action(\mathsf{us}, deter(1)), t_1))))$$
$$\mathbf{S}(\mathsf{us}, \mathsf{israel}, t_1, \mathbf{I}(\mathsf{us}, t_1(happens(action(\mathsf{us}, deter(1)), t_2))))$$
$$\mathbf{S}(\mathsf{us}, \mathsf{israel}, t_2, \mathbf{I}(\mathsf{us}, t_2(happens(action(\mathsf{us}, deter(1)), t_3))))$$
$$\mathbf{S}(\mathsf{us}, \mathsf{israel}, t_3, \mathbf{I}(\mathsf{us}, t_3(happens(action(\mathsf{us}, deter(1)), t_4))))$$

Instantiating the universal quantifier in $\boxed{\mathcal{A}_9}$ and applying $\Rightarrow$*-elim* and $\wedge$*-elim*, we end up with the following formulae; they represent the fact that Russia communicates the above statements to Iran.

$$\mathbf{S}(\mathsf{russia}, \mathsf{iran}, t_0, \mathbf{S}(\mathsf{us}, \mathsf{israel}, t_0, \mathbf{I}(\mathsf{us}, t_0(happens(action(\mathsf{us}, deter(1)), t_1)))))$$
$$\mathbf{S}(\mathsf{russia}, \mathsf{iran}, t_1, \mathbf{S}(\mathsf{us}, \mathsf{israel}, t_1, \mathbf{I}(\mathsf{us}, t_1(happens(action(\mathsf{us}, deter(1)), t_2)))))$$
$$\mathbf{S}(\mathsf{russia}, \mathsf{iran}, t_2, \mathbf{S}(\mathsf{us}, \mathsf{israel}, t_2, \mathbf{I}(\mathsf{us}, t_2(happens(action(\mathsf{us}, deter(1)), t_3)))))$$
$$\mathbf{S}(\mathsf{russia}, \mathsf{iran}, t_3, \mathbf{S}(\mathsf{us}, \mathsf{israel}, t_3, \mathbf{I}(\mathsf{us}, t_3(happens(action(\mathsf{us}, deter(1)), t_4)))))$$

Using the above block of formulae with universal instantiations of $\boxed{\mathcal{A}_{11}}$ gives us the following after performing $\Rightarrow$*-elim*:

$$\mathbf{B}(\mathsf{iran}, t_0, \mathbf{I}(\mathsf{us}, t_0, (happens(action(\mathsf{us}, deter(1)), t_1))))$$
$$\mathbf{B}(\mathsf{iran}, t_1, \mathbf{I}(\mathsf{us}, t_1, (happens(action(\mathsf{us}, deter(1)), t_2))))$$
$$\mathbf{B}(\mathsf{iran}, t_2, \mathbf{I}(\mathsf{us}, t_2, (happens(action(\mathsf{us}, deter(1)), t_3))))$$
$$\mathbf{B}(\mathsf{iran}, t_3, \mathbf{I}(\mathsf{us}, t_3, (happens(action(\mathsf{us}, deter(1)), t_4))))$$

Again, a simple instantiation of $\boxed{\mathcal{A}_{15}}$, along with $\Rightarrow$*-elim*, gives us:

$$\mathbf{B}(\mathsf{iran}, t_0, (happens(action(\mathsf{us}, deter(1)), t_1)))$$
$$\mathbf{B}(\mathsf{iran}, t_1, (happens(action(\mathsf{us}, deter(1)), t_2)))$$
$$\mathbf{B}(\mathsf{iran}, t_2, (happens(action(\mathsf{us}, deter(1)), t_3)))$$
$$\mathbf{B}(\mathsf{iran}, t_3, (happens(action(\mathsf{us}, deter(1)), t_4)))$$

The above formula states that Iran believes that the U.S. will impose economic sanctions on Iran. Let's assume reasonable values for Iran's estimates of the costs for enhancement (roughly on the order of the economic sanctions) and benefits (much more than the cost of enhancement):

$$\forall t : \mathsf{Moment}\ t < \mathsf{T}\ \mathbf{B}(\mathsf{iran}, t, cost(t) = 1 \wedge \Delta = 10)$$

Given that Iran has enough cognitive power for simple utilitarian calculations, we can derive:

$$\forall t : \mathsf{Moment}\ t < \mathsf{T} \Rightarrow \neg \mathbf{B}(\mathsf{iran}, t, cost_{sum}(t+1, \mathsf{T}) > \Delta - deter_{sum}(t+1, \mathsf{T}))$$

From the above formula, we can obtain a simple first-order derivation that shows that deterrence fails; an automatic proof of it is given in Figure 4. Figure 5 shows the fully machine-generated proof based on resolution and paramodulation in the SNARK automated theorem prover (see (Stickel, Waldinger, Lowry, Pressburger & Underwood 1994)).

3.1 On Automation

The semi-automated proof sketch presented above is for a small toy scenario serving to introduce our framework. For more detailed and realistic models, proof search might not be amenable to full automation. One of our solutions is to use a *library* of reasoning patterns that occur frequently in the domain of mindreading for nuclear deterrence. These common patterns can be captured as *methods* (or more frequently called $\lambda\mu$-methods) in a *denotational proof language* (DPL).[9] We use such methods to model mindreading in the false-belief task in (Arkoudas & Bringsjord 2009).

What would methods for the domain of nuclear deterrence look like? A subset of these methods would for instance be engineered for processing communications of a certain type released by either the U.S. or Iran. For example, if the U.S. releases a statement threatening economic sanctions, this set of methods could be fired to find out if that statement has had any influence on Iran's beliefs. Identifying common patterns of reasoning in this domain is a prerequisite for developing such methods, and this is work that has yet to be carried out in a systematic fashion. Once a sufficiently large base of such patterns is known, automating reasoning using methods is comparatively trivial.

3.1.1 Example: Automating False-belief Attribution

A common occurrence involving two or more agents is false-belief attribution. In false-belief attribution, an agent observes an another agent develop a false belief about some ϕ; the first agent knows what is actually the case (e.g. ϕ) but the second agent develops a belief that is not true ($\neg\phi$). The first agent then knows that the second agent has a false-belief. We have analyzed and modeled false-belief attribution in detail in (Arkoudas & Bringsjord 2009); others have as well, but to our knowledge no one else has used general-purpose methods instead of from-scratch representations for the particular task at hand. The specific task analyzed in this work is described here:

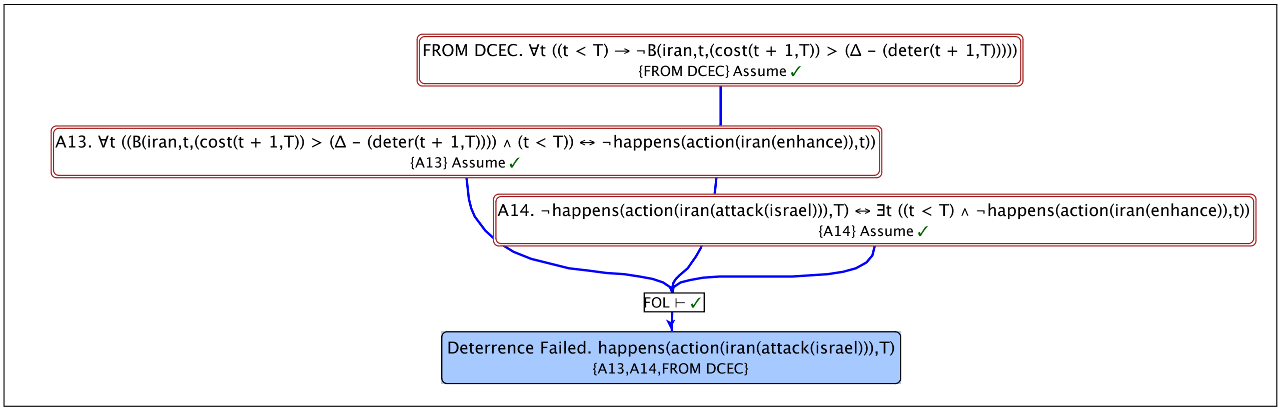
> "...a child (we will call her Alice) is presented with a story in which a character (we will call him Bob) places an object (say, a cookie) in a certain location l_1, say in a particular kitchen cabinet. Then Bob leaves, and during his absence someone else (say, Charlie) removes the object from its original location l_1 and puts it in a different location l_2 (say, a kitchen drawer). Alice is then asked to predict where Bob will look for the object when he gets back, ..."

The analysis centers around development of a micro-theory which predicts how agents with the capacity to model false-beliefs in others respond when probed with questions in a logically controlled natural language. More importantly, the above work includes development of $\lambda\mu$-methods to make the computation of false-belief attribution more tractable in a real-time modelling scenario.

We demonstrate a small scenario in which an automated system/human modeller analyzing state actors can attribute false beliefs. The scenario is illustrated in Figure 6.

[9] A DPL for a formal system such as $\mathcal{DCEC}^*$specifies a system for writing down proofs and abstractions of proof patterns. Please consult Arkoudas' (2000) dissertation for a comprehensive introduction to DPLs.

Figure 4: Slate Proof that Deterrence Will Fail Under Some Circumstances



FROM DCEC. ∀t ((t < T) → ¬B(iran,t,(cost(t + 1,T)) > (Δ − (deter(t + 1,T)))))
{FROM DCEC} Assume ✓

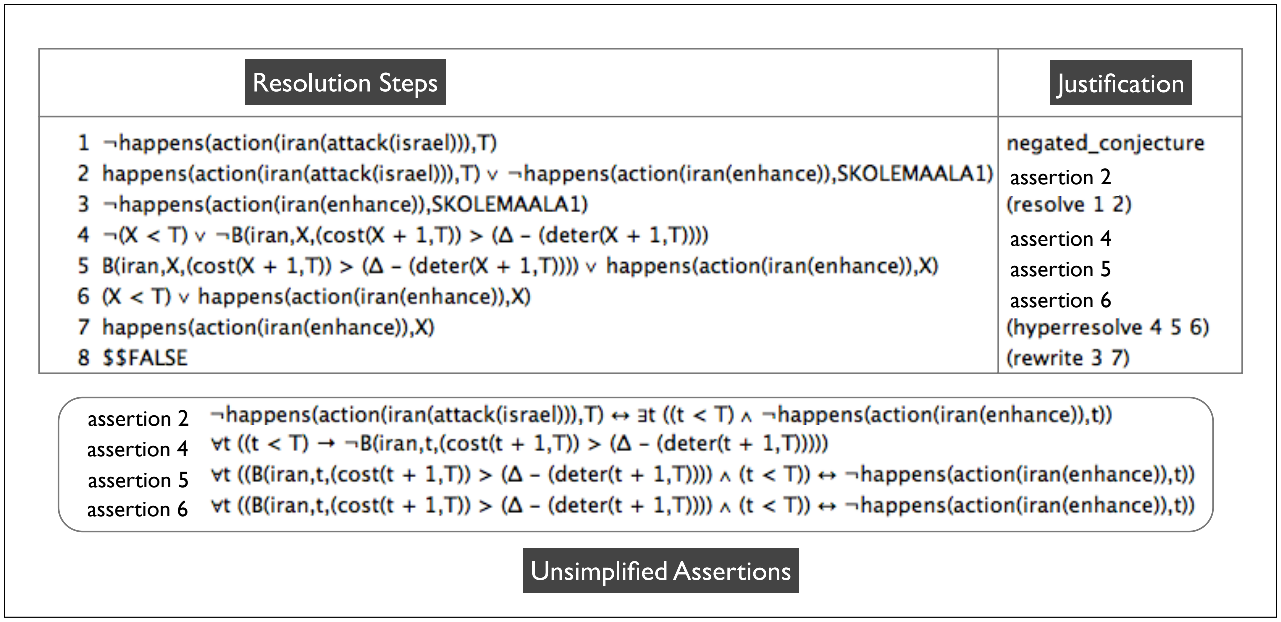
A13. ∀t ((B(iran,t,(cost(t + 1,T)) > (Δ − (deter(t + 1,T)))) ∧ (t < T)) ↔ ¬happens(action(iran(enhance)),t))
{A13} Assume ✓

A14. ¬happens(action(iran(attack(israel))),T) ↔ ∃t ((t < T) ∧ ¬happens(action(iran(enhance)),t))
{A14} Assume ✓

FOL ⊢ ✓

Deterrence Failed. happens(action(iran(attack(israel))),T)
{A13,A14,FROM DCEC}

Figure 5: Internal SNARK Proof (Automatically Generated)

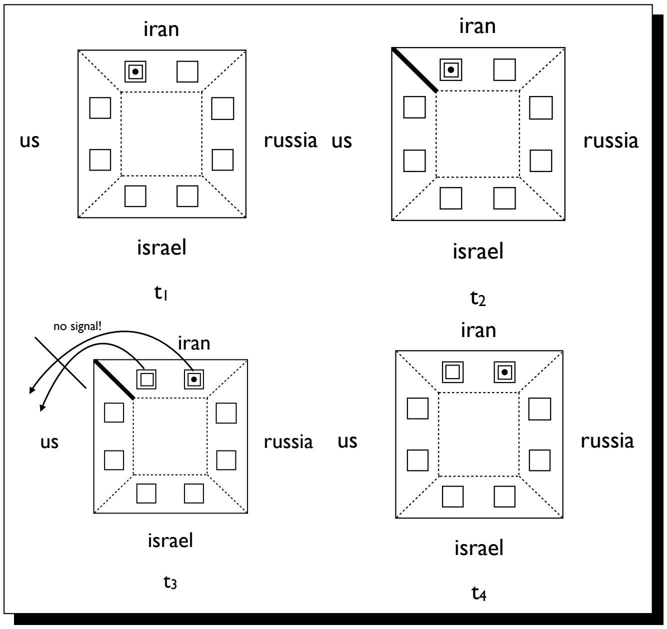


Resolution Steps	Justification
1 ¬happens(action(iran(attack(israel))),T)	negated_conjecture
2 happens(action(iran(attack(israel))),T) ∨ ¬happens(action(iran(enhance)),SKOLEMAALA1)	assertion 2
3 ¬happens(action(iran(enhance)),SKOLEMAALA1)	(resolve 1 2)
4 ¬(X < T) ∨ ¬B(iran,X,(cost(X + 1,T)) > (Δ − (deter(X + 1,T))))	assertion 4
5 B(iran,X,(cost(X + 1,T)) > (Δ − (deter(X + 1,T)))) ∨ happens(action(iran(enhance)),X)	assertion 5
6 (X < T) ∨ happens(action(iran(enhance)),X)	assertion 6
7 happens(action(iran(enhance)),X)	(hyperresolve 4 5 6)
8 $$FALSE	(rewrite 3 7)

assertion 2	¬happens(action(iran(attack(israel))),T) ↔ ∃t ((t < T) ∧ ¬happens(action(iran(enhance)),t))
assertion 4	∀t ((t < T) → ¬B(iran,t,(cost(t + 1,T)) > (Δ − (deter(t + 1,T)))))
assertion 5	∀t ((B(iran,t,(cost(t + 1,T)) > (Δ − (deter(t + 1,T)))) ∧ (t < T)) ↔ ¬happens(action(iran(enhance)),t))
assertion 6	∀t ((B(iran,t,(cost(t + 1,T)) > (Δ − (deter(t + 1,T)))) ∧ (t < T)) ↔ ¬happens(action(iran(enhance)),t))

Unsimplified Assertions

Figure 6: Modeling the Development of a False Belief



We have four time-points $\langle \mathsf{t}_1, \mathsf{t}_2, \mathsf{t}_3, \mathsf{t}_4 \rangle$. The figure illustrates visually/semantically all the information that is fed into the system. Each of $\{\mathsf{us}, \mathsf{iran}, \mathsf{israel}, \mathsf{russia}\}$ have two possible sites or locations in which strategic weapons can be built or located. A modeller studying this small micro-situation would need to show that the US possesses a false-belief based on certain publicly available information.[10] In the modeller's toolbox the methods $\langle M_1, M_2, M_3 \rangle$ shown below will be included. These methods and their derivation in a previous version of $\mathcal{DCEC}^*$ are discussed in Arkoudas & Bringsjord. The automated proof system comprised of methods for that task can be adapted to our micro-scenario.[11]

Input : A proposition which states that an agent a_1 perceives another agent a_2 performing an action.

$$\langle P_1 : \mathbf{P}(a_1, t, \mathit{happens}(\mathit{action}(a_2, \alpha), t)) \rangle$$

Output: A proposition which asserts that a_1 knows that a_2 knows that a_2 has performed this action.

$$\mathbf{K}(a_1, \mathbf{K}(a_2, t, \mathit{happens}(\mathit{action}(a_2, \alpha), t)))$$

Method 1: M_1

[10] In practice, one would not have access to all the information that would enable us to refute or deduce a certain statement with certainty. Incorporation of a philosophically sound mechanism to deal with uncertain information in $\mathcal{DCEC}^*$, ultimately necessary, is beyond the scope of the present introductory paper.

[11] There is a difference between the scenario discussed in Arkoudas & Bringsjord and ours here. In the former, the system/modeller has to derive the false-belief attribution statement of an agent it studies. This agent attributes a false-belief to another agent being studied: $\mathbf{K}(a_1, \mathbf{K}(a_2, \neg\phi))$. In our case, the system has to attribute a false-belief to an agent it is studying. This issue overlaps with that of representing first-, second-, and third-person *de se* statements correctly. We have addressed some of these issues in (Bringsjord & Govindarajulu 2013).

Input : A statement asserting that it is common knowledge event e initiates fluent f, a statement that a_1 knows that a_2 knows that e happens at t_1, a statement that $t_1 < t_2$ is common knowledge, and a statement that a_1 knows that a_2 knows nothing happens between t_1 and the later time t_2 that terminates e.

$$\begin{aligned}
&\langle P_1 : \mathbf{C}(t, \forall t' : \mathsf{Moment}\ \ \mathit{initiates}(e, f, t')),\\
&\ P_2 : \mathbf{K}(a_1, t, \mathbf{K}(a_2, t, \mathit{happens}(e, t_1))),\\
&\ P_3 : \mathbf{C}(t, t_1 < t_2),\\
&\ P_4 : \mathbf{K}(a_1, t, \mathbf{K}(a_2, t, \neg\exists e : \mathsf{Event}\ t' : \mathsf{Moment}\ \mathit{happens}(e, t') \land t_1 < t' < t_2 \land \mathit{terminates}(e, f, t')))\rangle
\end{aligned}$$

Output: a_1 knows that a_2 knows that the fluent f holds at t_2.

$$\mathbf{K}(a_1, t, \mathbf{K}(a_2, t, \mathit{holds}(f, t_2)))$$

Method 2: M_2

Input : A statement that a_1 knows that a_2 knows that the fluent f holds at t_2, a statement that $t_1 < t_2$ is common knowledge, and finally a statement that says that a_1 knows that there is no event e at any time between t_1 and t_2 which a_2 believes has happened and which a_2 believes terminates the fluent f.

$$\begin{aligned}
&\langle P_1 : \mathbf{K}(a_1, t, \mathbf{K}(a_2, t, \mathit{holds}(f, t_2)))\\
&\ P_2 : \mathbf{C}(t, t_1 < t_2),\\
&\ P_3 : \mathbf{B}(a_1, t, \neg\exists e : \mathsf{Event}\, t' : \mathsf{Moment}\ \mathbf{B}(a_2, \mathit{happens}(e, t')) \land \mathbf{B}(a_2, t_1, t', t_2) \land \mathbf{B}(a_2, \mathit{terminates}(e, f, t')))\rangle
\end{aligned}$$

Output: a_1 believes that a_2 believes that the fluent f holds at t_2.

$$\mathbf{B}(a_1, t, \mathbf{B}(a_2, t, \mathit{holds}(f, t_2)))$$

Method 3: M_3

4 Derived Desiderata

We are now in position to set out eight desiderata our framework satisfies, but which, as we shall soon see, extant frameworks are unable to, but should.[12] To ease exposition, we impose a four-part categorization on the eight desiderata. Please recall that we have not yet finished specifying the formal semantics for the intensional side of our framework (more on that below). Of course, for the extensional side, our framework inherits the formal model-theoretic semantics for first-order logic, into which any elements that might be wanted from decision theory and game theory can be easily cast. Here now are the requirements that must be satisfied by any promising framework in the domain of interest — requirements that our framework instantiates.

[12] One point about the list: It should be completely uncontroversial that any human-level extensional logic must subsume first-order logic. After all, standard treatments of the formalization of classical mathematics of necessity employ first-order logic (e.g., see Ebbinghaus, Flum & Thomas 1994). (And indeed such treatments, of necessity, employ more: at a minimum, the concept of an axiom schema that points to an infinite number of first-order formulas, as in the axioms of Peano Arithmetic.) Bringsjord is firmly of the opinion that (under the simplifying assumption that only finitary extensional logics are needed to model human mathematical practice) in fact *second*-order logic is needed and indeed is quite "cognitively natural," but for simplicity's sake here, $\mathcal{DCEC}^*$ is assumed to include, on the extensional side, only first-order logic. Actually, Bringsjord holds that human mathematical reasoning is *infinitary* in nature (which means that at a minimum "small" infinitary extensional logics (e.g., $\mathcal{L}_{\omega_1\omega}$, economically and elegantly summarized in Ebbinghaus et al. 1994) should be included in $\mathcal{DCEC}^*$, but for present purposes such logics are ignored. For brief discussion of cognitively realistic use of $\mathcal{L}_{\omega_1\omega}$, see (Bringsjord & Govindarajulu 2011).

Desiderata for a Promising Framework for MSP in Nuclear Strategy

1. **Expressive Formal Language, and Leveraging Thereof**. Our framework enables the modeling of interaction between multiple agents in the domain of nuclear strategy via formulae in a highly expressive formal language (which of course includes an explicit alphabet and grammar). Three immediate and specific benefits springing from this are:

   - **Complex Mindreading Formulae Enabled**. Our framework allows for the capture of such demanding sentences as: ($\star$) "Israel knows that Russia knows that the U.S. knows that Iran knows that Israel knows that Iran has said that Israel should be destroyed." (This is true in the real world.) This is just one arbitrary example, but it's one that, as we shall see when stacking game theory (in various forms) against our framework, is telling.

   - **Multiple Types of Intensional Operators**. Our framework offers the modeler a number of intensional operators "under one roof." Knowledge and belief are crucial in the nuclear-strategy domain — but so are perception, communication, intention, obligation, and so on. Our ultimate goal is to offer operators for *all* cognitive attitudes in play in the neurobiologically normal human mind operating within the domain of multi-agent nuclear strategy. Our foray into nuclear strategy is in this regard an instance of the general approach to modeling the minds of persons via logic-based techniques; the approach is described and defended in (Bringsjord 2008*a*).

   - **Importation of *Human-Level* Extensional Language**. It is well-known that human reasoning in the realm of mathematics (and the formal sciences generally), can only be formalized by extensional logics that subsume at least first-order logic (see footnote 12). $\mathcal{DCEC}^*$ includes first-order logic, as we have already pointed out.

2. **Proof-Energized**. Our framework is, as one might say, "proof-energized." This is concretized in three sub-requirements/desiderata:

   - **Proof Calculus**. Our framework, rooted as it is in $\mathcal{DCEC}^*$, enables the answers to queries to be not only answered, but to be *justified*. We saw this above: The queries there are answered, *and* a supporting proof is supplied as well.

   - **Proof Verification**. Our framework includes implemented algorithms for the verification of proofs supplied as justifications in conjunction with the answers returned in response to queries.

   - **Proof Discovery**. Our framework includes automated theorem proving technology, so that the machine can in some cases autonomously find the proofs that establish answers to queries. This is in large part what makes it possible for our framework to deliver predictions about the future states of groups of agents that interact around the issue of nuclear strategy. An automated proof, recall, is given in Figure 4.

3. **Full Computational Power of Logic-based AI**. Our framework claims and exploits the full power of computational techniques available to logic-based AI. Whereas game theory and decision theory and the like at most model relatively small parts of cognition/intelligence, our framework is a computational logic comprising everything the field of AI has to offer. For example, the full machinery of modern AI planning is available in our framework. We return to this point when discussing metagame theory. An overview and defense of logic-based AI is provided in (Bringsjord 2008*b*).

4. **Formal Semantics**. As we have admitted, the construction of our framework's comprehensive formal semantics isn't finished. However, our goal remains, and our framework, on the extensional side, inherits established schemes for formal semantics (e.g., classical model theory in the case of first-order logic). On the intensional side, as we explain below, we abandon model-theoretic and possible-worlds-style semantics in favor of a much more promising direction.

5 Game Theory and Extant Digital Games: Inadequate

5.1 Failure of Informal Game Theory

5.1.1 Informal Game Theory, Briefly

Informal game theory (IGT) is something that you are quite familiar with, or have at least to some degree studied and seen in action. Indeed, if before reading the present paper you have been but a casual student of (let alone a practitioner in) nuclear strategy, you have without question witnessed the application of IGT to such strategy (and indeed more on such application momentarily). IGT subsumes, for example, Nash-equilibrium theory, in connection with games of perfect and imperfect information, extensive games, finitely and infinitely repeated games, etc.[13] What you know as 'game theory' is in fact *informal* game theory (= IGT).

You may be inclined to protest, since what you have learned about game theory may in part be somewhat mathematical or — as you may proclaim — formal, at least. But such a protest is wrongheaded. The reason is that mathematics *itself*, from the standpoint of formal logic, is informal; mathematical logic, after all, was invented in order to formalize, and thereby clarify and demystify, standard mathematics and the practice thereof. The clarification came chiefly via three powerful and celebrated innovations: one, the invention of formal languages able to rigorize what had since at least Euclid been informal talk; two, a recognition that informal proofs are compelling because they can be expressed as formal proofs in at least one fixed proof calculus for first-order logic; and three, a recognition that theorems of the ordinary sort can in principle be derived via formal proof from a Turing-enumerable set of axioms (e.g., from ZFC set theory; see e.g. Ebbinghaus et al. 1994 for a succinct summary) expressed as formulae in the relevant formal language.[14] As is plain from the list of requirements satisfied by our new framework for modeling, simulating, and predicting events in the nuclear-strategy realm, these celebrated innovations are directly reflected in our own approach. And relative to these innovations, and what they meant for mathematics, standard game theory is undeniably informal. But the key question is: Is IGT inadequate for the nuclear-strategy domain? The correct answer is an affirmative one; we briefly defend it in the next section.

5.1.2 Why Informal Game Theory is Inadequate Relative to Our Framework

It's easy to see that IGT is inadequate relative to our desiderata. Recall that those desiderata are enumerated in Section 4. For convenience, while there are eight distinct requirements listed there, recall that we collapsed the first six under just two main headings. Under the stipulation that an approach which fails on any specific under either heading fails in fact under every requirement under that heading, we have a straightforward and rather poor "report card" for IGT in Table 3:

We conclude this section by pointing out that IGT has in fact been used to model multi-agent nuclear strategy — but unfortunately, this prior work plainly reveals the failure to meet our desiderata. For example, the 20^{th}-century arms race between the United States (**US**) and the Soviet Union (**SU**) has been rather anemically modeled as an instance of the well-known Prisoner's Dilemma (PD); see Table 4, which is taken

[13] Excellent and well-known introductory surveys include the following two: (Osborne & Rubinstein 1994, Osborne 2004). We make use of both shortly.

[14] Glymour (1992) has elegantly chronicled much of the intellectual history leading up to the first two of these three developments, a sizable share of the credit for which must go to Frege. For impressive work confirming the third (in connection with ZFC), see (Bourbaki 2004).

Table 3: IGT Measured by Our Desiderata

Desideratum	**Satisfied?**
Formal Language Present & Leveraged?	No
Proof-Energized?	No
Computational Power of Logic-based AI?	No
Formal Semantics?	No

directly from (Osborne 2004).[15] In this Table, the terms `Refrain` and `Build` denote the obvious actions. As is well-known, the action pair (`Build, Build`) is a (unique) Nash equilibrium here.[16] Whether or not this equilibrium corresponds to empirical results in the "real world" is an issue orthogonal to our fundamental observation.[17] As you can plainly see, this model, for instance, includes no explicit and formal reference to time, planning, communication, belief, knowledge, and so on; that is, the model fails to include the elements that are crucial to modeling, simulating, and predicting real-world actors in the nuclear-strategy realm.

SU **US**	Refrain	Build
Refrain	2 2	3 0
Build	0 3	1 1

Table 4: Arms Race Between the United States & Soviet Union as Prisoner's Dilemma

[15] Various other painfully simple games have been proposed as models of a two-player arms race; for example, the Stag Hunt (see e.g. Osborne 2004).

[16] For rusty readers: Denote an action profile by a^*; and by a^*_{-i} we mean the profile without a_i. Players are i and j; and u_i is a standard payoff function for player i's preferences. Then we define an action profile a^* to be a Nash equilibrium iff

$$\forall i \forall a_i (u_i(a^*) \geq u_i(a_i, a^*_{-i})).$$

[17] In point of fact, whether empirical results based on use of human subjects in behavioral experiments align with the Nash equilibrium in the case of PD at least *seems* to be an open question, since in no relevant work of the this kind is there a result that subjects conform to the Nash equilibrium at anything remotely close to 100%. In fact, it seems to be the case that the more subjects are allowed to engage in real-world communication, the less cognitively plausible the Nash equilibrium becomes. See, e.g. (Deutsch 1958, Rapoport, Guyer & Gordon 1976, Cooper, DeJong, Forsythe & Ross 1996). Under the reasonable assumption that communication between agents provokes cognition associated with the key intensional operators in our approach (e.g., *believes*), the fact that such communication reduces correspondence between the Nash equilibrium and empirical results can be taken as evidence in support of our approach.

5.2 Failure of Formal Game Theory

5.2.1 Formal Game Theory Encapsulated

What, then, about *formal* game theory (FGT)? We quite reasonably take FGT to to include a formal calculus $\mathcal{G} = \langle \mathcal{L}, \mathcal{P}, \mathcal{A} \rangle$ comprised of a well-defined language $\mathcal{L}$ for expressing statements about games in the form of axioms and theorems, and a proof calculus $\mathcal{P}$ by which to derive theorems from the axiomatization $\mathcal{A}$. The calculus $\mathcal{G}$ may also come with an interpretation I of $\mathcal{L}$ over some class of structures, but this is optional. Note that this understanding of a formal system is more closer to the idea that formal systems should deal with abstract *forms*. By this criterion, most of the traditional endeavors of not only game theory, but also domains as rigorous and precise as mathematics and physics, are not *form*al.

This state-of-affairs is slowly but surely changing. Very recently, mathematical and physical theories have been formalized in the above fashion to quite some extent in the multi-sorted "dialect" of first-order logic.[18] Such formalizations of mathematical and physical theories, in addition to helping us better understand the theories to which they have been applied, also provide us with an opportunity for formal verification of its theorems, and for automatic theorem proving (= discovery) — two of the desiderata on the above list for a truly effective framework for MSP in the nuclear-strategy realm. The most well-known of such formalizations is, of course, the axiomatization of Zermelo-Fraenkel set theory intended to serve as a foundation for mathematics (Potter 2004).[19] In the natural sciences, there has been some work devoted to formally modeling biological theories; for example, Woodger and Tarski, as early as 1937, carried out such work (Woodger & Tarski 1937). In the case of physics, extensive studies of relativity from an axiomatic point have been provided by Székely, Andréka, X. Madarász and Németi in (Madarász 2002, Andréka, Madarász & Németi 2002, Andréka, Madarász & Németi 2007, Székely 2009, Andréka, Madarász, Németi & Székely 2011), which is devoted to setting out the logical structure of relativistic theories in physics. We have recently utilized one such axiomatization for the special theory of relativity to prove a substantive theorem in physics: (Govindarajulu, Bringsjord & Taylor 2012).

5.2.2 Why Formal Game Theory is Inadequate

The most substantial formal calculus for game theory is that presented by Lorini and Moisan (2011). Though there have been studies of game theory from more cognitive and epistemic standpoints (e.g., Roy 2010 and Brandenburger 2008), these approaches do not furnish any formal calculus such as the one provided by us herein and by Lorini and Moisan (2011). At an absolute minimum, a formal calculus must have a well-defined syntax for specifying statements and a proof calculus for deriving theorems from axioms. Even though Lorini and Moisan (2011) provide a language for expressing extensive games, they fail to provide a full and formal proof calculus.

Another requirement for any formal game theory is that it should be able to express natural statements used in informal game theory transparently, without any encoding. Alas, the calculus provided by Lorini and Moisan (2011) falls short on this front too. To elaborate, suppose we have a statement of the following sort in a game that we are modeling:

> "At every time, player a knows what action player b has performed and communicates this to player c."

Any transparent formalization of this statement requires an object-level universal quantifier for expressing "At every time," epistemic operators for modeling "knows," and a modal operator for the speech act of

[18] Again, a nice treatment of multi-sorted in (Manzano 1996).

[19] We referred above to ZFC, but in point of fact Potter (2004) gives a variant. We leave such niceties aside.

declarative communication of propositions. The calculus by Lorini and Moisan (2011) fails in that it is based on *propositional* modal logic, which cannot express quantified statements such as the one above. Their calculus also does not have modal operators for expressing speech acts as simple as communication of declarative sentences between two agents. A full-fledged modal operator is necessary for expressing communication between agents, as what is communicated between agents is usually a proposition (even though the surface syntax used in such acts may be far from a simple proposition). Summarizing, existing axiomatizations of game theory fail on these dimensions of expressivity:

Extensional Dimension Should be able to express at the least first-order sentences with unbounded quantifiers.

Intensional Dimension Should be able to express at a minimum the most common of propositional attitudes (knowledge, belief, intention, desire etc.) and speech acts.

The formal calculus should have also have enough expressive power to separately model two kinds of statements:

Meta Level Statements that a game theorist would make. That is, statements *about* games.

Object Level Statements that a player playing the game would make. That is, statements *in* games.

The calculus needs to differentiate reasoning by agents in the game versus reasoning by the system or the game-theorist situated outside the game. Our simple way to model this is to have a system $\mathcal{L}_a$ for each agent or player a of interest and one $\mathcal{L}$ for the top-level theorist or system. Every derivation will then be tagged with its place of occurrence:

$$\Phi \vdash_a \phi$$

We point out that there can be subtle differences in the interpretation of statements depending upon subtle parameters in the reasoning in question. For example, indexicals such as now and I have coherent meanings only when used by agents situated in the game: when used by the system it's exceedingly difficult to pin down an exact meaning for such terms. We have made use of now, but not of the first-person pronoun. Lest the reader be tempted to regard our mention of indexicals to be merely ornamental, we acknowledge that in deeper, more fine-grained MSP in the nuclear-strategy realm it would be necessary to employ not only temporal indexicals, but I. For even if one agrees that treating nations as persons is a purely figurative move (in which case first-person beliefs would be ultimately eliminable in favor of talk of the beliefs of individuals), there will come a time when it will be necessary to model the first-person beliefs of world leaders whose cognition might well change the course of history. A tentative list of desiderata relating to indexicals, which any formal calculus of cognition should satisfy, can be found in (Bringsjord & Govindarajulu 2013).

To wrap up this section, note that Table 5 summarizes our analysis of how well FGT adheres to our standards.

5.3 Failure of Metagame Theory

5.3.1 What is Metagame Theory?

Metagame theory (MGT), invented and introduced by Howard (1971), supplements the concepts available in traditional game theory with an apparatus allowing decisions to factor in the *strategies* of players. MGT thus allows more complex and cognitively realistic modeling and simulation. Strategies model how players react

Table 5: FGT Measured by Our Desiderata

Desideratum	**Satisfied?**
Formal Language Present & Leveraged?	Yes, but insufficiently expressive
Proof-Energized?	No
Computational Power of Logic-based AI?	No
Formal Semantics?	Partially

to different types of opposing players; this alone makes MGT more promising as a framework for modeling nuclear strategy. For there can be little doubt that for instance the U.S. has, and must exploit, beliefs about strategies that Iran is likely to follow, versus those, say, Russia is likely to try. Another interesting aspect of MGT is that, if you will, it can be applied to itself. A level-1 metagame is applied only to a traditional game-theory scenario. In traditional game theory, as we saw above, a game simply revolves around making rational decisions in light of reasonable assumptions about possible outcomes and their dis/utility. The level-1 metagame revolves around choosing a level-0 (or traditional game-theory) strategy conditional on the level-0 strategies of the other players. A level-2 metagame works similarly, modeling rational level-1 choices based on the level-1 choices of other players. This nesting is what allows MGT to improve traditional game theory; and the improvement at least seems to be a promising one for the realm of nuclear strategy, as we have noted.

Consider the application of MGT to the classic example of the Prisoner's Dilemma (PD), in the simple cold-war form we visited above (Table 4). We have seen that the (unique) Nash equilibrium is action profile (`Build Build`). But of course the optimal action sequence is (`Refrain Refrain`). And indeed, in the real world, this is what we desire. (If we could return to a time when no player has built nuclear arms, and rational behavior for everyone is to refrain from building any nuclear weapons, and every player was rational, well, this would be a rather pleasant state-of-affairs.) MGT provides a framework in which we can prove the desired proposition that (`Refrain Refrain`) is the rational sequence for the arms-race version of PD. How? The outline of the proof is quite straightforward: We invoke the conjunction $\wedge$, negation $\neg$, and material conditional $\Rightarrow$ from elementary extensional logic, and deploy them to construct strategies. So, for instance, a level-1 strategy for **US** is:

$$\text{(L1)} \qquad \texttt{Refrain}_{\mathbf{SU}} \Rightarrow \texttt{Refrain}_{\mathbf{US}} \wedge \texttt{Build}_{\mathbf{SU}} \Rightarrow \texttt{Build}_{\mathbf{US}}$$

And this allows us to build a level-2 strategy for **SU**:

$$\text{(L2)} \qquad \text{(L1)} \Rightarrow \texttt{Refrain}_{\mathbf{SU}} \wedge \neg\text{(L1)} \Rightarrow \texttt{Build}_{\mathbf{SU}}$$

We can view pairs of strategies as directly analogous to action profiles in the standard game-theoretic framework, and can with Howard (1971) therefore define a **metaequilibrium** as the result in a standard game that corresponds to a Nash equilibrium in the metagame. This allows us to show that (`Refrain Refrain`) is rational, since it's a metaequilibrium; we have obtained our desired proposition.

5.3.2 Why Metagame Theory is Inadequate

Metagame theory is useful for modern political modeling. For example, Chavez & Zhang (2008) have elegantly applied MGT to the issue of Taiwan's independence from mainland China. Nonetheless, for MSP in the realm of nuclear strategy, MGT isn't competitive with our framework. To see this, we begin with a particular theorem T about MGT:

> For an n-player game, conditional strategies (Li) where $i > n$ have no impact on metaequilibria.

While T has been viewed by many as an advantage, it becomes a point of failure from the standpoint of iterated "mindreading" formulae. Why? Well, given that conditional strategies are believed, and of course that must be the case (indeed, that is the *point* of strategies: they are after all the targets of epistemic attitudes), T then places an unacceptably low limit on the order of beliefs about the strategies of others. There is no reason why in the real world an n-agent interaction cannot include $n+1$-, $n+2-$, $\ldots$ $n+m$-order conditional beliefs about the strategies of others — where these beliefs have real "bite." While the proponent of MGT could rebut that such beliefs would as a matter of mathematical necessity pertain to strategies expressed in ways outside MGT's regimented language for strategies in its narrow sense, this rebuttal is tantamount to admitting that the very concept of a strategy in MGT is itself quite narrow and limited relative to the real world. In our framework, strategies can be as elaborate as first-order logic, planning, and the event calculus allows.

Moving away now from formal criteria and looking at the objects studied by MGT, we note again for the record that MGT does more justice to what could actually happen in the real world if *bona fide* cognitive agents were to play such games. MGT could be thought of as providing a way of understanding the conditions and processes by which a real-world game comes to an equilibrium point or becomes stable. But alas, MGT fails to take this proper motivation toward full maturity. Even though MGT has some deeper cognitive notions than IGT and FGT, it does not capture anything doxastic or temporal. Rather, MGT captures only iterative strategies and outcomes. We illustrate our point with a small example.

Consider a game of three players (a,b,c) with the following action sets or strategy sets for this trio, respectively: $\{a_1,a_2\}$, $\{b_1,b_2\}$, and $\{c_1,c_2\}$. There can be many possible conditional strategies at different levels if the game is played by cognitive agents. If we adhere to a $\mathcal{DCEC}^*$-based view, the following three statements describe coherent higher-level strategies.

$S^1_{\mathcal{DCEC}^*}$: If player a does a_1 if a believes b will do b_2, then c intends to do c_1.

$S^2_{\mathcal{DCEC}^*}$: If player a has done a_1 if a knows b will do b_2, then c plans to do c_1.

$S^3_{\mathcal{DCEC}^*}$: If player a desires to do a_1 if a knows b will do b_2, then c plans to do c_1.

But these fine-grained strategies get collapsed into the following coarse strategy in MGT.

S^1_{MGT} : If player a does a_1 if b does b_2, then c does c_1.

Table 6 shows how MGT fares overall by our standards. In addition to failing to distinguish between different intensional concepts, such as belief, knowledge, intention, speaking/saying, etc., MGT also fails to address temporal notions inherent in games played in the real world. Temporal notions are intricately bound up with intensional statements. For example, consider *"If I believe that the other prisoner will defect, I will too"* vs *"If I believe that the other prisoner has defected, I will too."* These two statements have subtle differences in meaning, arising out of just small differences in tense, differences which cannot be ignored in any modeling framework. MGT's mathematical framework is completely agnostic even toward the existence of such intensional and temporal notions, yet such notions make modeling the interaction between the two agents in the two cases/statements non-trivial. The lack of the above features in MGT justifies our evaluation of how well MGT does on the semantic side.

Table 6: MGT Measured by Our Desiderata

Desideratum	Satisfied?
Formal Language Present & Leveraged?	No
Proof-Energized?	No
Computational Power of Logic-based AI?	No
Formal Semantics?	Partially

5.4 The Current Inadequacy of Digital and Tabletop Games

5.4.1 Relevant Digital and Tabletop Games, in Brief

“Would you like to play a game?”

This innocuous question is asked repeatedly by WOPR, the sentient artificial-intelligence system in the 1983 film *WarGames*. WOPR, short for ‘War Operation Plan Response,’ is designed to play and analyze strategy and war games in order to derive an “optimal” strategy for large-theatre combat, and one game it knows is Global Thermonuclear War. When a hacker breaks into the system and sets WOPR playing this game, a sequence of events is initiated which brings the U.S. military to DEFCON 1 and very nearly triggers World War III before, at the last moment, WOPR finishes its simulations and announces its conclusion. “A strange game,” it observes. “The only winning move is not to play.”

It is unsurprising, given the highly queasy nature of the subject, how few games there have been which deal with nuclear war in any genuine depth. Even for the open-minded and generally unafraid members of the games industry, it seems that this one subject, in which we include the nail-biting brinksmanship of *realpolitik* and international relations, is somehow unsavory, indecent almost. The idea of a game where one wrong step can — and often will — turn the world into a charred, radioactive cinder is not one which would have much appeal: As Tom Lehrer said when asked why he stopped writing satirical songs in the 1960s, “I was starting to feel like a resident of Pompeii being asked for a few humorous comments about lava.”

Nonetheless, there do exist some games which take up the subject of nuclear war. Numerous are text adventure games: notable examples include *Nuclear War Games*, in which the player must save the world from the actions of a rogue scientist, or the bleak but powerful *Ground Zero*, which casts the player as an Everyman figure who attempts to find ways to survive an imminent incoming nuclear strike. There are many games which include nuclear weapons as technologies for prosecuting war upon the player’s enemies, such as the *Sid Meier’s Civilizations* series and many tactical real-time and turn-based strategy games. We will examine neither of these types of games further here; rather, our focus is on games whose main gameplay revolves around tactics involved in executing, or, better yet, preventing, global thermonuclear war.

Nuclear War is a tabletop card game in which players represent nuclear powers. Initially released in 1965, it presented a satirical slant on the subject of nuclear war, an approach that has continued through subsequent releases at periodic intervals: the latest update, *Weapons of Mass Destruction*, appeared in 2004. With each addition, the gameplay has become more complex, though each unit is designed to be playable on its own, and each new game seeks to highlight the worries and attitudes of its period. At the beginning of a game, each player is dealt a number of ‘population cards,’ ranging in denomination from 1 to 25 million people. The object of the game is to protect this population, as its total loss results in elimination from the game. Various cards are dealt to the players in their turns: *secrets*, which steal or reduce enemy population; *propaganda*, which steal enemy population unless nuclear war has been declared; *delivery systems* and *war-*

heads, which combine to create weapons under the right circumstances; and *specials*, which can represent anti-missile systems or attack enhancements. After all secrets initially dealt have been played, the game continues in a state of 'Cold War' with cards being played one at a time in turn and resolved. If a warhead and a delivery system are available, the weapon *must* be used, initiating a 'Hot War' state: propaganda cards are now worthless and this state persists until one player is defeated, at which point 'Cold War' resumes. However, a player defeated by nuclear attack may launch a 'final retaliation' at the moment they leave the game, which often results in the defeat of another player, and ultimately a cascade of mutually assured destruction; if no players remain, there is no winner.

A more somber, yet arguably no less realistic, approach was taken by Chris Crawford in the 1985 game *Balance of Power*. This game was critically lauded at the time of its release and was quite popular, despite its brutally realistic, perhaps even cynical, depiction of political manoeuvring. In this game, the player takes the part of either the President of the United States or the General Secretary of the Communist Party of the USSR (and hence we think back to the PD-based arms-race model discussed in §5.1.2). The game lasts for eight turns, each turn representing one calendar year, and by the end of those years one must have maximized 'prestige' while avoiding global thermonuclear war. At the beginning of each year, a set of incidents and crises worldwide is presented and the player must respond to each one; responses include taking no action, sending a diplomatic note to the opponent, and undertaking military manoeuvres. Each response is met with a counter-response, which varies from backing down to escalation, and an opportunity for the player to respond again. Backing down, however, carries with it a loss of 'prestige,' which has major political repercussions, and is something both players seek to avoid; however, brinksmanship pushed too far can, and sometimes does, result in global war. If this occurs, the game ends instantly with a stark message displayed to the player.

The most recent game to approach the subject of global thermonuclear war is *DEFCON*. It is a difficult game to cite: its primary objective — destroy as much enemy population as possible while minimizing one's own casualties — is both simple and profoundly representative. A player, either human or AI, takes the role of the leader of one of six territories: North America, Latin America, Europe (excluding western Russia), Africa, Russia, and Asia. The game is played in real time, and as the game progresses, the DEFCON level steadily increases, changing the actions the player may undertake. At the beginning of the game, play is at DEFCON 5: players place units and fleets, later responding to new information as radar uncovers units within range. At DEFCON 3, conventional naval battle is authorized, but fleets may no longer be moved; and at DEFCON 1, the use of nuclear weapons is authorized. At this point the game proceeds until a certain percentage of all nuclear missiles has been launched; after which a victory countdown begins, at the end of which the final score is given. Scoring follows one of three schemes: *Default*, +2 points per million kills, -1 point per million casualties; *Survivor*, 1 point per million survivors in the player's territory; or *Genocide*, 1 point per million kills. Alliances can be forged, broken, and renegotiated between human players (CPU-player alliances may only be formed at the start of the game); given that only player victories are possible, this frequently leads to the breakdown of alliances and a total free-for-all before the end of the game.

5.4.2 Why Such Games are Inadequate

A game, when considered as an entity in itself, is a carefully crafted artifact, and rarely more so than when it seeks to represent or encapsulate a concept that is large in scope. There is a delicate balance to be drawn between *realism* and *playability*: too little of the former and the game will not be a fair representation; too little of the latter and it becomes perceived as a chore. It is reasonably safe to say that the greater the scope of the game, the more difficult this balancing act becomes; that it is not impossible is evidenced by games such as the *Europa Universalis* series, in which players struggle for supremacy in a historically accurate model of Middle-Ages Europe. All games, however, must compromise in their design, leaving out certain aspects and promoting others: there is, after all, only so much space and so much complexity even the most hardened of players will tolerate. In the examples cited in the previous section, it is easy to see that the designers omit large sections of the nuclear game in order to focus on what they consider key. *Nuclear War* takes a satirical approach to Cold War politics and MAD, and omits the finer points of *realpolitik* and strategy; *Balance of Power* is strongly focused on politics and brinksmanship while leaving war simulation well alone; *DEFCON* is primarily strategic and almost blackly humorous in its stark first-principles assertion that "everybody dies."

The fundamental problem, then, is one of complexity. In the game of nuclear strategy, which includes the subgame of nuclear deterrence, there are almost infinitely many variables, some known and some unknown, connected in a myriad of ways and strongly interdependent. We may be able to see clearly the existence of many of these variables and connections; the existence of some of them, we might be able to infer accurately; others we can only guess at, and still others we must concede to be mysterious: we do not know if they exist or not, and, in either event, whether their existence — or even their non-existence — change the state or nature of the game. Like no other, the "nuclear game" is one of imperfect information and hence heavy on belief but shorter on knowledge, and bluff and counterbluff. Aspects of these phenomena are credibly simulated in the tabletop arena, in games such as *Battlestar Galactica*, where there are enemy Cylon agents passing as humans among the human crew struggling to survive, or *The Resistance*, where players may be Imperial Agents or brave Resistance members fighting for freedom; in both cases, guesswork and deduction play pivotal roles in the players' shaping the outcome of the game. But the bottom line is that the full range of phenomena in the real-life nuclear strategy appears never to have been captured in either digital or tabletop games.

This is not to say that we are the first group to appreciate the recalcitrance of nuclear strategy. The German-born economist Oskar Morgenstern, co-developer with John von Neumann of game theory, famously likened the Cold War to poker at a time when the underlying strategy used by strategists was derived from a different kind of game entirely. He (1961) wrote: "The cold war is sometimes compared to a giant chess game between ourselves and the Soviet Union [...]. The analogy, however, is quite false, for while chess is a formidable game of almost unbelievable complexity, it lacks salient features of the political and military struggles with which it is compared." And, elaborating:

> Chess is, to begin with, a game of complete information. That is, the chess opponent has no unknown cards, no means at his disposal which the other player cannot see and know all about. Every move is made in the open; consequently (AND THIS IS MOST IMPORTANT), there is no possibility of bluffing, no opportunity to deceive. Obviously, these conditions are far removed from political reality, where threats abound, where the threatening nation has to weigh the cost not only to its enemies, but to itself, where deceit is certainly not unheard of, and where chance intervenes, suddenly favoring first one side, then another. [...] The present cold war situation makes this need for strategic perception not only apparent but imperative. Thermonuclear disaster might be triggered at any time by a few false steps which become increasingly difficult to avoid as new conflict zones, like Cuba and Congo, arise. Furthermore, nuclear weapons are spreading ominously to more nations while the ability to deliver them anywhere, from any point on earth, is already in the hands of the two superpowers. [...] With bluffs so much easier to make and threats so much more portentous than at any previous time in history, it is essential not only for our own State Department but for the entire world to understand what bluffs and threats mean; when they are appropriate; whether they should be avoided at all costs; in short, that is the sanest way to play this deadly, real life version of poker. (Morgenstern 1961, p. SM14)

The 'way' that Morgenstern alludes to, we claim, aligns with what our framework offers. The real-life game that we are in is one that appears to be without visible end; furthermore, to extend the poker analogy, there also appears no way to leave the table. We cannot take heed of WOPR's sage advice by making "the only winning move" and declining to play: the only obvious course of action is continue to raise and re-raise our stakes, hoping that our initial wager, based on our best MSP tools at the time, will indeed be good enough. As we have seen, we cannot rely upon game theory to come to our rescue: that stalwart of the decision sciences, despite originally having been conceived as an attempt to explore the game of poker in mathematical terms (McManus 2009), has the unfortunate flaw of assuming that all agents are thin one-dimensional utility maximizers rather than robust cognizers who, like the players in the high-stakes real-word nuclear game, believe, know, and communicate.

6 Objections; Rebuttals

6.1 "But the core computational challenges are Turing-uncomputable!"

We imagine some skeptics responding as follows: "In your approach, prediction is proof-driven, which means that in order to determine whether some initial situation S would lead to an outcome O, one must model S via a set Γ_S of formulae, model O via a formula ϕ_O, and then decide whether ϕ_O is provable from Γ_S. But since you are by design well beyond first-order logic (after all, even your 'causal core' of the event calculus employs full-blown FOL), this decision-problem is Turing-undecidable. Getting a standard computing machine to automatically decide the question is therefore impossible. Getting an intelligence analyst to do it is also unrealistic, since with all due respect to such folks, they don't find writing simple SQL queries (which are of course at bottom simple FOL formulae) trivial. So who or what will discover the derivation and how? And, this question assumes that there is a derivation to be found. What if there isn't? How will someone know that O cannot arise from S; that is, that $\Gamma_S \nvdash \phi_O$?"

Numerous rebuttals disarm this objection; we give three:

1. Part of the objection is in the form of "proof by cases," in that two possible mechanisms for proof construction are considered, and in each case ruled out. The two mechanisms are: computing machine operating autonomously (as that after all is the only mode related to undecidability theorems for theoremhood in whatever proof calculus is under consideration), and intelligence analyst. But this is alas an instance of the fallacy *false dilemma*. Even if these two routes are assumed for the sake of argument to be dead ends, there is a third avenue, and it's one that parallels the history of first-rate analysis in the realm of nuclear strategy: namely, employ brilliant and formal-methods-trained professionals whose working lives are devoted to the specific subject matter at hand. Ironically, the semi-automated approach that we have presented points directly down this promising road. Hypothetically, we ourselves could work as a team for half a century, focusing for perhaps the first two decades on modeling and simulation with no more than 12 "agents." If human beings working essentially in solitude can routinely discover proofs that settle questions much more syntactically complicated than the ones at the heart of those that motivate our proposed new approach, it begs the question against us to assume that a group like ours couldn't discover proofs that could then be communicated to, and discussed with, the relevant analysts, in natural language. To encapsulate the rebuttal, if the present objection's dichotomy is valid, then our planet's need for program verification can't possibly be met, since (i) program verification is Turing-unsolvable (it's at least Π_2), and (ii) garden-variety programmers (let alone naïve users of programs) are unable to produce proofs of a program's correctness.

2. The objection given here, if telling against our particular research and development, would in one blow sweep away *all* r&d based on automated theorem proving. Program verification is just one additional

example from countless ones. For example, the attempt to build machines capable of human-level mathematics and logic should be immediately abandoned if the present objection is sound. The reason is that such activity in the human sphere comprises conjecture generation, proof discovery, proof generation, and then finally proof checking and verification — all in a Turing-uncomputable context. The famous case of Fermat's Last Theorem provides a perfect case in point. The solutions provided by Wiles (1995, 1995) establish a proof of a theorem within a general case that is Turing-uncomputable. Just as the likes of Wiles can seek answers to problems that are in a Turing-uncomputable space of problems, we can systematically seek answers to problems that are likewise in a general Turing-uncomputable space.

3. As is well-known, the Turing-undecidability of theoremhood in FOL (and in logics that subsume FOL, such as our $\mathcal{DCEC}^*$) is commonly established by reduction of the decision problem in question to the halting problem (HP) (e.g., see Ebbinghaus et al. 1994). This mathematical fact provides a very practical guide as to how to proceed in the case of our approach to nuclear strategy: viz., if an answer isn't returned after an intolerably long period of time, just switch off the CPU and default to the verdict that no answer is ever going to to be returned. There is ultimately no practical problem in our approach whatsoever, in light of this policy, which is a standard one in the "real world," commonly used even in the automatic checking and grading of student-submitted programs in introductory programming. In real life, definitive counter-models are simply not needed for many applications; and they are certainly not needed in AI. If they *were* needed, then AI researchers and engineers whose approach is the "flip side" of ours (e.g., researchers who have no proof theory and no proof discovery/verification theory and technology whatsovever, but rather only a model-based approach, e.g., Fagin, Halpern, Moses & Vardi 2004) would be instantly known to be fundamentally misguided. Just as one cannot rationally declare that such researchers are dead in the water because they have no proof theory, one cannot delcare that our approach is untenable because we have no model-theoretic machinery beyond that for first-order logic.[20] And, as we explain below, it's a *non sequitur* to infer from the absence of a model-theoretic/possible-worlds semantics that no formal semantics is available.

6.2 "But you don't have a formal semantics!"

Here, a critic says: "By your own admission, your framework doesn't yet include a formal semantics. There may be some virtue in your confession, but the fact remains that absent a formal semantics, your framework is profoundly problematic."

Since we aspire to eventually provide a formal semantics that covers every singly relevant part of our framework, it follows deductively, given minimal assumptions about our level of rationality as logicist AI researchers interested in building applications in the realm of nuclear strategy, that we regard our here-reported work to be incomplete. But in no way does the absence of a formal semantics compromise the tools we can already provide to the community of military strategists and researchers interested is modeling, simulating, and predicting in this realm. There are myriad reasons why this is so; we encapsulate three of them now. After giving these reasons, we conclude the section with a gentle warning to readers that the formal semantics we are developing for both epistemic and deontic operators (and indeed ultimately for *all* intensional operators) is in no way in what we see as the failed tradition of standard model-theoretic semantics and it's naïve extension in the possible-worlds direction).

1. Scientists have been engaging in modeling, simulation, and prediction in the domain of nuclear strat-

[20] Reminder: We do appropriate the model theory for FOL for $\mathcal{DCEC}^*$formulae that are purely extensional.

egy for over half of the 20th century, and the dominant techniques that first took root at the inception of this work (at the RAND Corporation) have now been updated and sustained in our new century — and targeted at a set of nation states that form a superset of the state actors that we featured above in our sample analysis. *All of this work is stunningly informal; as such, not only is there no formal semantics, but there is no formal alphabet, no formal grammar, no proof theory, no rigorous algorithms, no computer programs, no automation . . . nothing whatsoever of the sort.* Even without a fully developed formal semantics, our framework is hence rather an advance, and it puts those who employ it ahead of the standard, primitive game.

Quick and decisive confirmation of the brute, empirical fact that standard work not only lacks a formal semantics, but lacks the rigor provided by formal languages, proof theories, and proof verification/falsification, is provided by the very recent, rather ambitious RAND monograph from Delpech (2012): *Nuclear Deterrence in the 21st Century*. Delpech rejects use of *any* formalism to better analyze and predict in the realm of nuclear strategy. For example, even standard game theory (IGT in our earlier analysis) is rejected as sterile formalism:

> For some, game theory was the recognizable sign of serious strategists, their ability to formalize situations. For others, it was ignorance of strategy, history, culture. The leading experts in game theory . . . sought strategic rules . . . and emotions were sidelined. . . . Strategists like Kohn showed little interest in game theory, and the complexity of the contemporary world does not encourage thinkers today to show any more interest than he did. We now believe that we must learn more about regional issues and know specifically who the opponents are in order to make meaningful policy, rather than turn them into abstract "players" in some heuristic game of questionable relevance to the real world. (Delpech 2012, 48–49)

Our framework, as has been seen, is explicitly able to model some of the rather important real-world cognitive aspects of human beings brusquely left aside by game-theoretic formalisms: knowledge, belief, perception, and communication, for instance. And of course our framework allows humans and machines to verify that reasoning is valid. In short, if semantics-less approaches expressed in casual and ambiguous natural language are taken seriously in the real of nuclear strategy, then the absence of a full semantics in our case cannot be allowed to obscure the virtues of the preliminary version of our formal framework.

An important related rebuttal neutralizes the concern that we're overly sanguine about the ability of formal modelers in collaboration with analysts to build sensible models. The reason is that this concern, if truly worrisome, let alone fatal, would immediately vitiate the work carried out by the Delpechs of the world, and by the many politicians, diplomats, and military strategists who by the very nature of their jobs must assemble informal sensible models. Our approach is at the highest level the time-honored and tried-and-true transfer from the informal to the formal-and-computational, and the gain thereby of the added power that comes from rendering things in clear and digital form suitable for exploiting the incredible and ever-growing speed of high-speed computers.

2. The objection here also runs afoul of the fact that the frameworks with which we are competing (canvassed, recall, in §5), for example game theory, are themselves bereft of a formal semantics — but that is arguably the least of their problems. How could our new paradigm possibly be untenable because its semantic side is unfinished, given that the other candidate paradigms for modeling, simulation, and prediction in the nuclear-strategy realm lack *any* formal semantics, *and* lack, on the syntactic/proof-theoretic side, our ability to model epistemic and deontic operators over the full event calculus in precise, machine-checkable fashion? You will for instance search in vain through the

afore-cited (Osborne & Rubinstein 1994) for any formal semantics, and any axiomatization. And even in simple modeling challenges where game theory is thought to excel, it can be easily shown that the representational machinery provided is painfully informal and inexpressive when compared to the cognitive event calculus (e.g., see our formal modeling of the famous-in-game-theory chain-store paradox: Bringsjord, Govindarajulu, Eberbach & Yang 2012). In short, the frameworks with which we are in competition don't even have their syntactic and proof-theoretic sides developed, let alone the semantic side. We provide a formal signature for all modeling and simulation, as well as an axiomatization in each case. This level of rigor and reach has yet to be achieved in other formalisms traditionally associated with modeling, simulation, and prediction in the nuclear-strategy realm (indeed, in *any* realm). We are therefore rather ahead of the game, despite the fact that our framework's semantics is unfinished.

3. It's important to note that mathematics itself has no finished formal semantics. The only way that the work of mathematicians in formalizing some domain, and then proceeding to prove theorems in that domain under a certain set of assumptions (a kind of work that is a perfect match, in general, for what our framework enables, as seen above: we provide a formalization, make certain assumptions, and proceed to prove theorems, with help from the machine), could have a formal semantics would be if the work in question is firmly within a standard first-order language, a standard proof theory for first-order logic, and standard model-theoretic semantics. But there are many proofs and theorems and mathematics that cannot be directly expressed (as a matter of mathematics itself) in this humble machinery.[21] The absence of a formal semantics doesn't stop mathematicians and logicians from charging ahead, and the fact that our formal semantics is at this point unfinished shouldn't prevent us from making progress.

And now, as promised, we conclude this section with the promised gentle warning.

We fear that some readers will assume that our formal semantics for $\mathcal{DCEC}^*$must take some sort of denotational form; for instance, a form in line with standard model theory, or the natural extension of standard model theory to cover modal logics. This assumption would be entirely erroneous. While for purely extensional logics and formal theories built therefrom, for instance $\mathcal{EC}$ as a standalone theory not targeted by intensional attitudes, we happily appropriate standard truth-condition approaches, but we completely and utterly reject all such approaches for intensional operators and any logics that include such operators, in favor of what we see as a much more sensible and promising approach to formal semantics for intensional concepts: *proof-theoretic semantics*, seminally introduced by Gentzen (1935*b*).[22] We regard it to be nothing short of a travesty that possible-worlds semantics, which is a naïve extension of model theory for extensional logic, have been awkwardly applied to the case of intensional operators. In $\mathcal{DCEC}^*$, the meaning of any intensional operator, or more precisely the meaning of any formula based on such operators, consists in the proof- and argument-theoretic context around the relevant constituents of that formula. To say that some person knows ϕ is for us to say that the corresponding **K**-led formula κ is situated within a hypergraph that represents the proof of κ, where that proof reflects the cognitive relationship between — as Kreisel (1971) points out when reflecting on proof-theoretic semantics — a cognizer and her reason-based cognition.

[21] Mathematicians and logicians routinely make use of both second-order constructs and in fact expressions that are themselves infinitely long (which is part of the main reason that infinitary logics arrived on the scene in the first place); see (Karp 1964) for a litany of theorems that when directly represented require infinitary logic. Note that it's not even possible to express finitude in first-order logic, in this (i.e., standard FOL with standard model-theoretic semantics) simple machinery. That is, there is no set Φ of first-order formulae such that a (standard model-theoretic) interpretation I satisfies all $\phi \in \Phi$ if and only if that interpretation is finite.

[22] For the English version, see (Gentzen 1935*a*).

Bringsjord's believing that you, reader, are educated, has precious little to do with metaphysical meanderings like possible worlds and "ways the world could have been" and so on, but everything to do with the fact that he either has in mind, or is poised to bring it to mind, that since all readers of a paper such as this one are educated, and you are a such a reader, it follows proof-theoretically that you are educated. Proof-theoretic semantics is now exceptionally mature and promising (in no small part because Prawitz carried Gentzen's torch forward; e.g., see Prawitz 1972), and has even been quite nicely applied to the problem of providing a semantics for natural language (e.g., see Francez & Dyckhoff 2010). Please note, finally, that the integrity, coherence, and power of proof-theoretic semantics implies that the following claim is highly controversial: Traditional meta-properties grounded in standard old-style model theory (e.g., soundness) must be the yardstick for gauging even extensional logics. *A fortiori*, our intensional approach must not be unfairly trammeled by the application of such a yardstick.

7 Conclusion

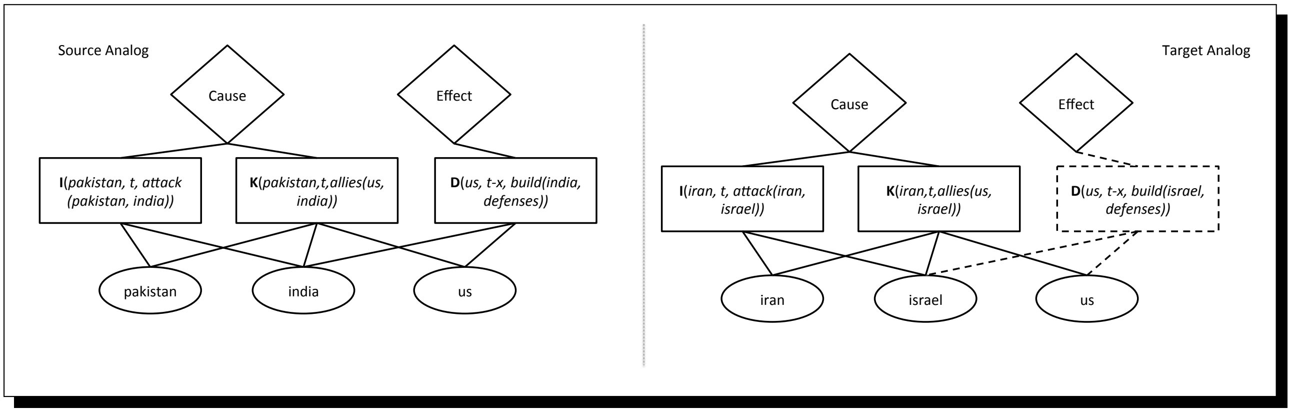
The foregoing is admittedly but a quick summary of the first phase of our lab's foray into at-once rigorous-and-cognitively-detailed modeling and simulation in the realm of nuclear strategy, so as to enable reliable prediction. Our ultimate goal is to provide those in US Defense and Intelligence, and civilian decision-makers with whom they work (as well as their counterparts among the allies of the United States), a logico-computational framework that will allow possible futures to be seen, and the desired ones to be secured. Relative to this goal, which needless to say is not unambitious, we have of course at best conveyed the gist of our technical foundation, and methodology. The immediate next steps to advance our framework are five in number: (1) fine-tune the four-agent model described above in order to use the framework to issue some genuinely actionable predictions with respect to Iran, each under some particular set of interesting assumptions; (2) advance the proof-theoretic formal semantics for the intensional side of our framework; and (3) create and implement a digital game that coincides with our framework, and therefore offers humans the opportunity to both learn the framework and, at least to a degree, plumb the future while playing the corresponding game. In addition, we stand ready to evaluate competing frameworks outside of the game-theoretic tradition — but at present we are unaware of any such competitors, despite (4) ongoing search.[23]

Furthermore, and this is step number (5), we are preparing to extend our framework by allowing for inferences that marry deductive and non-deductive techniques. We fully realize that domain experts leverage their knowledge of real-world scenarios from the past, in order to understand the present, and work toward predicting the future. It would for example be manifestly imprudent for anyone to use our framework in the real world without thorough knowledge of things like the Cuban Missile Crisis, which provide data from which to reason by analogy to the pressing problems of the present. More specifically, **analogico-deductive reasoning**, our novel integration of hypothetico-deductive and analogical reasoning, allows for propositional content to be analogically inferred from existing knowledge. This inferred knowledge could then be further subjected to deductive analysis in order to evaluate its consistency and plausibility. This approach, explored (in domains other than mind reading) in (Bringsjord & Licato 2012, Licato, Govindarajulu, Bringsjord, Pomeranz & Gittelson 2013), can be easily applied to the type of reasoning we have discussed in this paper. For example, we may have historical knowledge of the three-agent "nuclear" relationship between Pakistan, the United States, and India. From this we could analogically infer the sort of knowledge pictured in Figure 7, although we must note here that this is a very simplified example. Our framework could then attempt

[23] So-called "BDI" logics in AI are much less powerful than our framework (for commentary and analysis, see e.g. Arkoudas & Bringsjord 2009), so our search has already covered them, and classified them as non-competitive. To here prevent the present paper from heading toward the length of a monograph, we do not recapitulate, let alone expand, our evaluation of BDI systems.

to determine whether this new knowledge is inconsistent with knowledge acquired through other means, enriching the predictive power of our framework as a whole.

Figure 7: Example Source and Target Analog. The items in dotted lines are inferred analogically.


Source Analog
Cause
Effect
I(pakistan, t, attack (pakistan, india))
K(pakistan,t,allies(us, india))
D(us, t-x, build(india, defenses))
pakistan
india
us
Target Analog
Cause
Effect
I(iran, t, attack(iran, israel))
K(iran,t,allies(us, israel))
D(us, t-x, build(israel, defenses))
iran
israel
us

Above all, we are steadfastly committed to the formalization of deliberative, multi-agent mindreading of arbitrary complexity (in symbiosis with deontic, temporal, and causal reasoning). Such formalization we take to be a *sine qua non* for the achievement of our ultimate goal, and while we have miles to go before we can sleep soundly despite the spectre of nuclear war in an asymmetrical world that makes the old-century simplicity of the Cold War seem a paradise by comparison, this formalization has herein been, we hope, informatively introduced.

References

Andréka, H., Madarász, J. X. & Németi, I. (2002), *On the Logical Structure of Relativity Theories*, Alfréd Rényi Institute of Mathematics, Budapest, Hungary. With contributions from A. Andai, G. Sági, I. Sain, and Cs. Töke. http://www.math-inst.hu/pub/algebraic-logic/olsort.html.

Andréka, H., Madarász, J. X. & Németi, I. (2007), Logic of Space-Time and Relativity Theory, *in* M. Aiello, I. Pratt-Hartmann & J. van Benthem, eds, ‘Handbook of Spatial Logics’, Springer, pp. 607–711.

Andréka, H., Madarász, J. X., Németi, I. & Székely, G. (2011), ‘A Logic Road From Special Relativity to General Relativity’, *Synthese* pp. 1–17.
URL: *http://dx.doi.org/10.1007/s11229-011-9914-8*

Arkoudas, K. (2000), Denotational Proof Languages, PhD thesis, MIT.

Arkoudas, K. & Bringsjord, S. (2008*a*), Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, *in* T.-B. Ho & Z.-H. Zhou, eds, ‘Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)’, number 5351 *in* ‘Lecture Notes in Artificial Intelligence (LNAI)’, Springer-Verlag, pp. 17–29.
URL: *http://kryten.mm.rpi.edu/KA_SB_PRICAI08_AI_off.pdf*

Arkoudas, K. & Bringsjord, S. (2008*b*), ‘Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task’, *PRICAI 2008: Trends in Artificial Intelligence* pp. 17–29.

Arkoudas, K. & Bringsjord, S. (2009), 'Propositional Attitudes and Causation', *International Journal of Software and Informatics* **3**(1), 47–65.
URL: *http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_041709.pdf*

Artic Computing, Ltd., Brandesburton, UK (1984), 'Ground Zero', Game; see URL.
URL: *http://retrobrothers.hubpages.com/hub/GroundZero*

Bello, P., Bignoli, P. & Cassimatis, N. (2007), Attention and Association Explain the Emergence of Reasoning About False Belief in Young Children, *in* 'Proceedings of the 8th International Conference on Cognitive Modeling', University of Michigan, Ann Arbor, MI, pp. 169–174.

Bourbaki, N. (2004), *Elements of Mathematics: Theory of Sets*, Verlag, New York, NY. This is a recent release. The original publication date was 1939.

Brandenburger, A. (2008), 'Epistemic Game Theory: Complete Information', *The New Palgrave Dictionary of Economics, Eds. SN Durlauf and LE Blume, Palgrave Macmillan* .

Bringsjord, S. (2008*a*), Declarative/Logic-Based Cognitive Modeling, *in* R. Sun, ed., 'The Handbook of Computational Psychology', Cambridge University Press, Cambridge, UK, pp. 127–169.
URL: *http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf*

Bringsjord, S. (2008*b*), 'The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself', *Journal of Applied Logic* **6**(4), 502–525.
URL: *http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf*

Bringsjord, S. & Govindarajulu, N. S. (2011), 'In Defense of the Unprovability of the Church-Turing Thesis', *International Journal of Unconventional Computing* **6**, 353–374.

Bringsjord, S. & Govindarajulu, N. S. (2013), 'Toward a Modern Geography of Minds, Machines, and Math', **5**, 151–165.
URL: *http://www.springerlink.com/content/hg712w4l23523xw5*

Bringsjord, S., Govindarajulu, N. S., Eberbach, E. & Yang, Y. (2012), 'Perhaps the Rigorous Modeling of Economic Phenomena Requires Hypercomputation', *International Journal of Unconventional Computing* **8**(1), 3–32.
URL: *http://kryten.mm.rpi.edu/SB_NSG_EE_YY_28-9-2010.pdf*

Bringsjord, S. & Licato, J. (2012), Psychometric Artificial General Intelligence: The Piaget-MacGyver Room, *in* P. Wang & B. Goertzel, eds, 'Theoretical Foundations of Artificial General Intelligence', Atlantis Press.
URL: *http://kryten.mm.rpi.edu/Bringsjord_Licato_PAGI_071512.pdf*

Chavez, A. & Zhang, J. (2008), Metagame Strategies of Nation-States, with Application to Cross-Strait Relations, *in* H. Liu, J. Salerno & M. Young, eds, 'Social Computing, Behavioral Modeling, and Prediction', Springer US, New York, NY, pp. 229–238.
URL: *http://dx.doi.org/10.1007/978-0-387-77672-9_25*

Chris Crawford; Mindscape, Boulogne-Billancourt, France (1985), 'Balance of Power', Game; see URL.
URL: *http://en.wikipedia.org/wiki/Balance_of_Power_(video_game)*

Cooper, R., DeJong, D., Forsythe, R. & Ross, T. (1996), 'Coöperation Without Reputation: Experimental Evidence from Prisoner's Dilemma Games', *Games and Economic Behavior* **12**, 187–218.

Delpech, T. (2012), *Nuclear Deterrence in the 21st Century: Lessons from the Cold War for a New Era of Strategic Piracy*, RAND Coporation, Santa Monica, CA. ISBN/EAN: 9780833059307.

Deutsch, M. (1958), 'Trust and Suspicion', *Journal of Conflict Resolution* **2**, 265–279.

Don Eskridge (2009), 'The Resistance', Game; see URL.
URL: *http://en.wikipedia.org/wiki/The_Resistance_(party_game)*

Ebbinghaus, H. D., Flum, J. & Thomas, W. (1994), *Mathematical Logic (second edition)*, Springer-Verlag, New York, NY.

Fagin, R., Halpern, J., Moses, Y. & Vardi, M. (2004), *Reasoning About Knowledge*, MIT Press, Cambridge, MA.

Fantasy Flight Games, Roseville, MN (2008), 'Battlestar Galactica: The Board Game', Game; see URL.
URL: *http://www.fantasyflightgames.com/edge_minisite_sec.asp?eidm=18&esem=1*

Flying Buffalo Games, Scottsdale, AZ (1965, 1983 and subs.), 'Nuclear War', Game; see URL.
URL: *http://en.wikipedia.org/wiki/Nuclear_War_(card_game)*

Francez, N. & Dyckhoff, R. (2010), 'Proof-theoretic Semantics for a Natural Language Fragment', **33**, 447–477.

Gentzen, G. (1935*a*), Investigations into Logical Deduction, *in* M. E. Szabo, ed., 'The Collected Papers of Gerhard Gentzen', North-Holland, Amsterday, The Netherlands, pp. 68–131. This is an English version of the well-known 1935 German version.

Gentzen, G. (1935*b*), 'Untersuchungen über das logische Schlieben I', *Mathematische Zeitschrift* **39**, 176–210.

Glymour, C. (1992), *Thinking Things Through*, MIT Press, Cambridge, MA.

Goble, L. (2003), 'Preference Semantics for Deontic Logic. Part I: Simple Models', *Logique et Analyse* **46**, 183–184.

Govindarajulu, N. S., Bringsjord, S. & Taylor, J. (2012), 'Proof Verification and Proof Discovery for Relativity'. Presented at the First Conference on Logic and Relativity 2012, Budapest, Hungary.

Horty, J. (2012), *Reasons as Defaults*, Oxford University Press, Oxford, UK.

Howard, N. (1971), *Paradoxes of Rationality: Theory of Metagames and Political Behavior*, MIT Press, Cambridge, MA.
URL: *https://mitpress.mit.edu/books/paradoxes-rationality*

Introversion Software; London, UK. (2006), 'DEFCON: Everybody Dies', Game: see URL.
URL: *http://www.introversion.co.uk/defcon/*

Jaśkowski, S. (1934), 'On the Rules of Suppositions in Formal Logic', *Studia Logica* **1**, 5–32.

Karp, C. (1964), *Languages with Expressions of Infinite Length*, North-Holland, Amsterdam, The Netherlands.

Kreisel, G. (1971), A Survey of Proof Theory II, *in* J. E. Renstad, ed., 'Proceedings of the Second Scandinavian Logic Symposium', North-Holland, Amsterdam, The Netherlands, pp. 109–170.

Licato, J., Govindarajulu, N. S., Bringsjord, S., Pomeranz, M. & Gittelson, L. (2013), 'Analogico-deductive Generation of Gödel's First Incompleteness Theorem from the Liar Paradox', *Proceedings of the 23rd Annual International Joint Conference on Artificial Intelligence (IJCAI–13)* .

Lorini, E. & Moisan, F. (2011), 'An Epistemic Logic of Extensive Games', *Electronic Notes in Theoretical Computer Science* **278**, 245–260.

Madarász, J. X. (2002), Logic and Relativity (in the light of definability theory), PhD thesis, Eötvös Loránd University, Budapest, Hungary.

Manzano, M. (1996), *Extensions of First Order Logic*, Cambridge University Press, Cambridge, UK.

McManus, J. (2009), *Cowboys Full: The Story of Poker*, Macmillan, New York, NY.

McNamara, P. (2010), Deontic logic, *in* E. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2010 edn. The section of the article discussing a dyadic system is available at: http://plato.stanford.edu/entries/logic-deontic/chisholm.html.

Morgenstern, O. (1961), 'The Cold War is Cold Poker', *New York Times Magazine* (February 5), SM14. This appeared in the Magazine section.

Mueller, E. (2006), *Commonsense reasoning*, Morgan Kaufmann.

Osborne, M. (2004), *An Introduction to Game Theory*, Oxford University Press, Oxford, UK.

Osborne, M. & Rubinstein, A. (1994), *A Course in Game Theory*, MIT Press, Cambridge, MA.

Potter, M. (2004), *Set Theory and its Philosophy: A Critical Introduction*, Oxford University Press, Oxford, UK.

Prawitz, D. (1972), The Philosophical Position of Proof Theory, *in* R. E. Olson & A. M. Paul, eds, 'Contemporary Philosophy in Scandinavia', Johns Hopkins Press, Baltimore, MD, p. 123134.

Rapoport, A., Guyer, M. & Gordon, D. (1976), *The 2 x 2 Game*, University of Michigan Press, Ann Arbor, MI.

Roy, O. (2010), 'Epistemic Logic and the Foundations of Decision and Game Theory', *Journal of the Indian Council of Philosophical Research* **27**(2), 283–314.

Softgold, Melbourne, Australia (1985), 'Nuclear War Games', Game.

Stickel, M., Waldinger, R., Lowry, M., Pressburger, T. & Underwood, I. (1994), Deductive Composition of Astronomical Software From Subroutine Libraries, *in* 'Proceedings of the Twelfth International Conference on Automated Deduction (CADE–12)', Nancy, France, pp. 341–355. SNARK can be obtained at the url provided here.
URL: *http://www.ai.sri.com/∼stickel/snark.html*

Székely, G. (2009), First-Order Logic Investigation of Relativity Theory with an Emphasis on Accelerated Observers), PhD thesis, Eötvös Loránd University, Budapest, Hungary.

Wiles, A. (1995), ‘Modular Elliptic Curves and Fermat’s Last Theorem’, *Annals of Mathematics* **141**(3), 443–551.

Wiles, A. & Taylor, R. (1995), ‘Ring-Theoretic Properties of Certain Hecke Algebras’, *Annals of Mathematics* **141**(3), 553–572.

Woodger, J. H. & Tarski, A. (1937), *The Axiomatic Method in Biology*, Cambridge University Press, Cambridge, England.

Wooldridge, M. (2009), *An Introduction to Multiagent Systems*, 2 edn, Wiley, New York, NY.