Robotic Misdirection, For Good Causes

Strategically Deceptive Reasoning in Artificial Generally Intelligent Agents

Max Fowler
fowlml01@students
.ipfw.edu

Aaron Thieme
thieac01@students
.ipfw.edu

John Licato
jlicato@ipfw.edu

Analogical Constructivism and
Reasoning Lab
Department of Computer
Science
Indiana University-Purdue
University Fort Wayne

ABSTRACT

Deception is a core component of human interaction and reasoning, and despite its negative connotation, it can be used in positive ways. We present our formalization behind strategic deception, one such potentially positive form of deception. We use the Cognitive Event Calculus ($\mathcal{CEC}$) to model strategic deception, building on prior formalizations. First, we provide a brief overview of deception's definitions within existing literature. Following this discussion, $\mathcal{CEC}$ is described and we present $\mathcal{CEC}$-style inference rules for strategic deception. These rules and a positive motivating deception example are used to show how we can solve the problem of strategic deception. This proof is demonstrated both through application of our rules and by adapting our rules for MATR (Machina Arachne Tree-based Reasoner) to show how proving can be performed by automatic reasoners. Finally, we discuss what future steps can be taken with strategic deception.

Keywords

Artificial General Intelligence; AI; Deception; Automatic Prover; Cognitive Event Calculus

1. INTRODUCTION

One ultimate goal of Artificial General Intelligence (AGI) is to finally bridge the gap between man and machine and create systems capable of human level thought and reasoning. Waser's work aiming to clarifying AGI as a field postures that the positive goal of AGI is that human style reasoning systems will be universal problem solvers for the world [23]. Some approaches to AGI take a formalized mathematical basis, such as Hutter's AIXI agent used to model artificial death by Martin et. al. [14]. Others take the approach that we should develop computational logics which provide reasoning strong enough to model human level reasoning and hopefully not see us all be killed, as Bringsjord argued [3]. This paper takes the latter approach, offering a formalization of a wonderfully human action - *strategic deception*.

Lying and deceiving are quintessential elements of human reasoning and interaction. Hippel and Trivers consider deception, and specifically the co-evolution between deceivers and those who are deceived, to be a major contributing factor to the evolution of human intelligence [22]. This makes having a formalization for deception ideal, such that we may better understand our own cognitive systems. Further, an understanding of deception opens the kinds of interactions we can model for the field of artificial general intelligence. A greater wealth of interactions will hopefully allow for more advances in the field.

Deception is often considered negative (e.g. lying to one's wife about a mistress, deceiving one's boss about work accomplished, tax evasion), yet deception does have positive benefits. Many of these benefits exist in the field of creating artificially intelligent systems to assist humans. Sakama describes a medical assistant that may not always tell patients the truth, much like doctors must sometimes practice deception in their bedside manner to keep patients calm [18]. Another medical example includes a diagnosis robot. Assume there is a minuscule chance of a patient having lupus and treating lupus will kill that particular patient for some reason if that patient does *not* have lupus. It would be ideal, then, for a medical diagnosis robot to not inform the doctor about the small chance of the disease being lupus until other options are exhausted.

It is further reasonable to think of cases where a deceptive artificial agent can provide more security than other agents, to the benefit of humans. Consider the case of an artificial generally intelligent robot guarding a school's research lab. The robot has a key to access the lab, knows all the members of the lab, and is instructed to avoid conflict when dealing with potential intrusions into the lab. A student of ill morals approaches the robot, intent on gaining entry to the lab by lying about being a lab member's friend. Logically, it would be well within the robot's rights to tell the individual to leave. However, this goes against the directive to avoid conflict, as a rude response could result in the would-be thief becoming desperate, violent, or more scheming in response. We wish to give our robot agent the ability to deceive the thief into believing the robot is unable to help them directly by lying that it does not have the lab key anymore. This provides a safer, more diplomatic diffusion of the situation. In what manner, then, can we teach our agents how to deceive, like a human could, to avoid this conflict?

Deception is well agreed upon as requiring success to be

called such [12]. Lying is generally accepted as requiring the statement of a belief that is false to the speaker [13]. These agreements serve as a cornerstone for the formalization of deception but are unsatisfying in their abstractness. Other researchers have attempted to define specific requirements for deception and lies to function. Forbus argues that deception necessarily assaults agents' predictive abilities and argues for an analogical reasoning approach towards understanding the mechanism of deception [8]. Stokke argues for the assertion model of lying and claim that assertions should be used to create common ground. This common ground provides a shared set of beliefs between agents that is pivotal for lying to proceed [20]. Chisholm and Feehan add that lies necessitate that the liar wish for their lie to be believed by another [7].

Multiple formalizations exist for various forms of deception and deceptive situations. Sakama creates a general formalization of deception, based on van Ditsmarch's formalization of lying; Sakama calls this the *agent announcement framework* [18, 21]. The work provides a solid backbone for formalizing general deception but can be notationally unintuitive. Licato's work showed how the modal logic based cognitive event calculus ($\mathcal{CEC}$) can be used to elegantly model the nested layers of beliefs required to perform the deception shown in the show *Breaking Bad*, in a fashion that lends itself well to automatic reasoners [9]. There exists room to marry the efficiency in modeling provided by $\mathcal{CEC}$ with rules designed specifically to formalize deception, similar to the work of Sakama and van Ditsmarch.

We present a $\mathcal{CEC}$ formalization for deception while defining strategic deception. First, we will present the definition of deception and strategic deception we will use in this paper. Then, we define the problem of strategic deception: what is necessary for strategic deception, why it is useful, and what the success and failure conditions are. Following, we formalize our reasoning approach by expanding upon $\mathcal{CEC}$ with new inference rules. As an aside, we develop forms of Sakama's deception rules, translating from the *agent announcement framework* to $\mathcal{CEC}$. Finally, we show how by using $\mathcal{CEC}$ and MATR (Machina Arachne Tree-based Reasoner), an automatic reasoning system, an artificial generally intelligent agent can reason over the lab guarding situation and successfully diffuse the issue.

2. DEFINING DECEPTION AND STRATEGIC DECEPTION

Before defining strategic deception, we must present the definition of general deception this paper uses. The OED defines deception by saying that it is "to cause to believe what is false" [1]. Mahon's work rejects this as too simple, as it allows for mistaken deception and inadvertent deception [13]. Mistaken deception concerns cases where an agent leads another to believe a false formula that the agent itself believes. A recent example of inadvertent deception can be found in the striped dress which led the internet to debate, "Is this dress white and gold or black and blue?" [17]. Mahon's presents a traditional definition of deception, D1, that requires deception to be an intentional act: "To deceive $==_{df}$ to intentionally cause to have a false belief that is known or believed to be false" [13]. We prefer to align ourselves with Mahon's D2, though, as it restricts the deception to only cases where the deceiver causes the deception, rather than a third party or outside force: "A person x deceives another person y if and only if x causes y to believe p, where p is false and x does not believe that p is true." This definition is most in agreement with Sakama's definition of deception, which is part of what we will use to define strategic deception [18]. By default, this definition does not require a lack of truthfulness, which means one can deceive by telling the truth. This definition also has no requirement for making statements, which means non-verbal communication and even non-communication, such as placing a briefcase in a room, can be used to deceive.

We define strategic deception as a specialized form of Chisholm and Feehan's *positive deception simpliciter*, the form of deception in which one agent contributes to another acquiring a belief [7]. In strategic deception, the deceiving agent must want something of another agent: generally, to act upon or in-line with the deceiver's goal. This goal can be in a negative form (e.g. I do not want this agent to eat my sandwich). We define a strategically deceptive agent as follows:

> (SD) An agent a is strategically deceptive to another agent b IFF agent a causes b to believe ϕ, where ϕ is false and a believes that ϕ is false, by causing b to believe some false statement ψ, selected such that believing ψ requires b to develop belief in ϕ, using some strategy to accomplish an overall goal γ.

In order for an agent to be deceptive in a general sense, there are a number of conditions that must be met. Sakama agrees with the common contention that deception, by definition, requires success [18]. We include this in our definition of strategic deception. Castelfranchi's earlier work on deception requires that the *addressee* believes the *speaker* is attempting to benefit or assist them, and thus be trustworthy, and believe that the agent is not ignorant [6]. Further, McLeod's summarized definition of trustworthiness requires vulnerability on the part of the *addressee*, requires some assumed competence on the part of the *speaker*, and requires that the *addressee* think well of the *speaker* within some context [16]. For this paper, we assume trust is given unless a deception is caught, as the establishment of trust is not within our scope.

Deception functions differently in regards to different kinds of agents. Sakama's formalization of deception primarily focuses on credulous agents, which are defined in the *agent announcement framework* as agents who believe the speaker is sincere [18, 21]. We consider it unlikely that deceiving credulous agents is worth investigating, as such agents are bound by their nature to adopt any belief directed at them. For our purposes, we are more concerned with the *agent announcement framework's* skeptical agent: in brief, skeptical agents are belief consistent agents, only adding beliefs to their belief set that are consistent. We refer to Sakama's skeptics as *maximally belief consistent agents*, to avoid confusing them with other definitions of skepticism.

Strategic deception requires that our agent lie. That is, agent a must believe some statement ϕ, yet act as if they believe $\neg\phi$. The *agent announcement framework*'s set up for a lie based deception requires that the *listener* come to believe a false statement ψ, based on the idea of believing the *speaker* is truthful. That is, ϕ justifies belief in ψ to agent b. We adopt the directionality that $\psi\ justifies\ \phi$. This more

naturally opens up the ways our agents can lie. For example, while it is possible that a can literally say ϕ implies ψ, lies by omission are desirable. Consider the case of eating a coworker's sandwich and being accused after the act. Saying, "The fact that I am a vegetarian means Bob, and not I, must have eaten your ham sandwich," may convince the accuser. However, just saying "I'm a vegetarian," implies that one could not have eaten a ham sandwich. Further, saying, "I'm a vegetarian, but I saw Bob near the fridge earlier," accomplishes the same thing at the first sentence without directly lying. If a lie is by omission, it is not involved in the dialogue and may be harder to pick up on. This makes our overall deception hard for agent b to check, which is one of the conditions put forward for the successful selection of lies by Forbus [8]. In the sandwich example, if we never mention the possibility of eating the sandwich at all, agent b may simply not think about that possibility and blame Bob instead. This certainly holds true for maximally belief consistent agents, in the event Bob eating the ham sandwich is a reasonable explanation: we remove the alternative that we ate the sandwich completely.

In Section 3, we discuss how we know strategic deception has succeeded and how strategies for deception are designed. In Section 4 and onward, we discuss our formalization and how we use rules to prove deception.

3. HOW WE KNOW WE HAVE STRATEGICALLY DECEIVED

Strategic deception requires the creation of a strategy. This strategy is made up of the statements agent a can make in order to deceive agent b. In order to form a strategy, we must know the *domain* of our situation. More specifically, a must know the domain they are using to deceive b. The domain includes a, b, and any other entities who may be related to this particular school lab or the lab's parent department. It further includes beliefs a has about these traits and beliefs a believes b has. For our original example, some domain beliefs are believing the department has a secretary, believing that secretary helps students, and believing that r helps students and secretaries. From the domain, then, we create a strategy consisting of our goal γ, a ψ generated to justify our lie, and any supporting μ statements we wish to use.

Strategic deception necessitates the generation of a false statement ψ by the *speaker*. The selection of an appropriate ψ is a difficult quandary. We do not make an effort to rate specific ψ against each other within the same domain directly. Instead, we concern ourselves only with ensuring a ψ is an appropriate choice. To determine if a false statement ψ is appropriate for a given deception, we consider the set $\mathfrak{P} = p_1, ..., p_i$ of all beliefs related to the situation agent b holds. A simple heuristic, then, allows us to rapidly rule out candidate ψs.

> ψ removal heuristic 1: if $\mathfrak{P} \cup \{\psi\} \vdash q$ for an arbitrary q, ψ is unfit to choose as the false statement justification for our deceptive agent's lie due to being contradictory to b's beliefs.

Finalizing the selection of which ψ an agent decides to say is more difficult that ruling out bad ψ. A good ψ must help advance the deceiving agent's goal. That is, belief that ψ justifies ϕ should lead to an agent acting upon the deceiver's goal γ. This leads to a second heuristic for ψ selection.

> ψ removal heuristic 2: if the chosen ψ does not lead to the deceived agent acting upon γ, then the ψ is unfit to choose for justification as its selection does not lead to success.

As a final consideration for ψ generation, we need to consider how a ψ is actually formed. Without bounding what information an agent can use to generate a ψ, we risk allowing an agent too much information that may not be relevant to the problem at hand, which may bog the decision making process down significantly. Therefore, we require that a given ψ be chosen only if it is within the domain of the strategic deception being carried out. This domain includes traits about the situation, such as the location and agents involved, as well as the *speaker*'s beliefs and beliefs about the *addressee*'s beliefs. An example of a domain is defined along with our proof further on.

This same domain is useful for the generation or recruitment of supporting μ_ns. All supporting statements must lend credence to ψ and must belong to the same domain as ψ. Officially, this means that any given μ is selected in order to make ψ believable to an *addressee*, and thus allow for deception to proceed. For the belief consistent agents we use, this is sufficiently handled by requiring any chosen μ makes ψ belief consistent with the *addressee*'s belief set. Recruitment of μ_ns can be carried out by a adopting beliefs it thinks b has. Generation, meanwhile, should merge facts traits from the domain with either beliefs a has or a believes b has or with lies or bluffs that are consistent with b's beliefs. To provide a set heuristic for ruling out μ options, we use:

> μ removal heuristic: if the chosen μ does not help lead to the deceived agent believing ψ, then the μ is unfit to choose as the justification as it does not lead to success.

One way to consider μ in a general sense is to consider μ's relevance. In that respect, the above heuristic can be summed up as the relevance of *the belief in* μ in regards to *the belief in* ψ.

Strategic deception also requires an established mechanism for asserting beliefs and establishing common grounds. Stokke contends that lying requires some assertion from *speaker* to *addressee* [20]. We address this in our inference rules later on using $\mathcal{CEC}$ $\mathbf{S}$ operator. We wish to point out that here, we operate with $\mathbf{S}$ in the linguistic sense of stating sentences. It is sufficient for words to be used in the communication, but they can be spoken or written. Non-verbal addressing is acceptable for general deception and assertion, as supported by Chisholm's formative work [7]. Stokke further mandates that common ground between *speaker* and *addressee* are required for deception to succeed [20]. We agree, as this is consistent with Sakama's belief consistent agents. This is further consistent with requiring belief consistent agents be made to believe the *speaker* believes what they are asserting for deception to succeed [18]. This is why later on, we require agent a to not only make agent b believe the lies but also make agent b believe that agent a believes the lies as well.

It is important to consider the success and failure conditions for strategic deception in some detail. As most work on deception requires, we require strategic deception to be successful. Further, as strategic deception is goal motivated,

a's goal must be met. Strategic deception fails, then, in the following situations:

1. A given $\neg\phi$ or $\psi \rightarrow \neg\phi$ fails to be consistent with b's beliefs and b rejects a's trustworthiness as a result, believing they are being lied to
2. A failure of the deception due to irrationality on b's behalf
3. A failure of the strategy used if b is successfully deceived yet does not act in the way a intends

Case (1) is a clear cut failure of deception. Case (2) is trickier. We define irrationality on agent b's behalf as agent b rejecting, rather than adopting, a belief consistent belief they are exposed to. This still means the deception fails, and thus agent a was not deceptive. However, we wish to make clear that agent a's *strategies* do not fail in (2). An agent practicing perfect strategic deception can always fail through no fault of their own in the event of (2) occurring. Finally, case (3) is interesting in that it is a failure not of the deception, but of the *strategy* used. The strategy is defined as the selection of ψ and supporting μs, as well as any other steps taken during the strategic deception process. If agent b comes to be deceived, yet does not act as agent a intends (or does not act at all), agent a has failed. This makes strategic deception potentially more flimsy than general deception, as agent b's inaction results in a's failure.

4. FORMALIZING LIES AND DECEPTION IN $\mathcal{CEC}$

We begin by describing $\mathcal{CEC}$. Arkoudas and Bringsjords' cognitive event calculus ($\mathcal{CEC}$) is a first-order modal logic framework that expands upon Kowalski's event calculus [4, 10]. The event calculus itself is a first-order logic with types. It features *actions*, or *events*, to represent actions that occur. *Fluents* are used to represent values which can change over time and can be propositional or numerical in nature. Time is represented with *timepoints* which can be either continuous or discrete. In summary, the event calculus is used to model how *events* affect *fluents* through *time*, allowing for the modeling of event chains [19].

The event calculus models these event chains through the acts of clipping and starting fluents through events. If a fluent exists and has not been clipped (ended or stopped by an action) at a time t, then it is said the fluent holds at t. For any time t, a fluent will hold for that time so long as it has yet to be clipped. Events, then, are responsible for both initiating fluents and clipping them. An event chain can trace how a fluent is effected by the events occurring to it.

$\mathcal{CEC}$ creates an event calculus for cognition. It uses modal operators for belief ($\mathbf{B}$), knowledge ($\mathbf{K}$), and intent ($\mathbf{I}$). $\mathcal{CEC}$ avoids possible-world semantics, in favor of a more computationally reasonable proof-theoretical approach. An attempt is made to model natural deduction as closely as possible, to best represent human-style reasoning [15]. Two of the most important departures $\mathcal{CEC}$ has are as follows:

- $\mathcal{CEC}$'s inference rules and logical operators are restricted to the contexts for which they are defined, to prevent problems that can occur with overreaching rules.
- Underlying inferences use constantly refined inference rules. This is used instead of cognitively implausible strategies, despite the latter having some potential use.

The $\mathcal{CEC}$ formulae tend to include an agent, a time, and a nested formula. When agent a believes ϕ at time t, we write $\mathbf{B}(a, t, \phi)$. Similar syntax is used to say an agent perceives, ($\mathbf{P}$),knows ($\mathbf{K}$), an agent says something ($\mathbf{S}$). There are some special operators that do not follow this trend. $\mathbf{C}$ is used to establish a common belief, while $\mathbf{S}$ has a directed syntax for agent a to declare a formula to agent b. Intention is handled as an intent to perform an action. While an agent can intend to act at time t, the intention identifies a time t' when that intention will be acted on. $\mathcal{CEC}$ uses *happens* as an operator to launch an action [5].

$\mathcal{CEC}$ also addresses the idea of agents being *able* to perform actions, using affordances. Affordances are actions an agent can perform starting at a time t. All possible actions an agent can take are the agent's affordance set. We say $isAffordance(action(a, \phi), t)$ when at time t, and beyond agent a can perform that action. This was added to $\mathcal{CEC}$to allow belief creation to be handled on an afforded basis, rather than on an immediate basis following logical closure [11]. As a further trait of $\mathcal{CEC}$, actions tend to require the *happens* operator. For example, $happens(a, t, act(\phi))$ means that at time t, it happens that a has performed the *act* action on some formula ϕ. If a instead *intends* to perform that action, the following syntax is used: $happens(a, t, intends(a, t, act(\phi)))$.

In the rules below, we introduce a *supports* operator. This operator conveys that the first argument causes the second to become believable. For a *maximally belief consistent agent*, if μ supports ψ, then that simply means that μ is consistent with b's beliefs and then allows ψ to be consistent. Much like $justifies$, there is room to grow *supports* for different kinds of agents, in regards to relatedness and similar factors, that is not addresses in this paper's scope.

Moving on, we set out to model deception in $\mathcal{CEC}$. We start our formalization by converting some of Sakama's deception axioms to $\mathcal{CEC}$. We leave most of the nuances of the Sakama's framework out of this paper, though we do walk our readers through two of Sakama's axioms. First, consider Sakama's A2, the axiom covering a liar's understanding of their having lied:

$$(A2)[¡_a\phi]\mathbf{B}_a\psi \equiv \mathbf{B}_a\neg\phi \supset \mathbf{B}_a[¡_a\phi]\psi \tag{1}$$

The *agent announcement framework*, while concise, can be difficult to expand. The left hand side says that after a's lying announcement of ϕ, agent a believes ψ. The right hand side of the equivalence is the implication that if agent a believes $\neg\phi$, then agent a believes that after their lying announcement of ϕ, ψ is true. The essential component of this rule, in regards to the modal $\mathcal{CEC}$, is that when agent a lies about ϕ, they believe ψ is true. One problem here is the implicit assumption that $\neg\phi$ leads to ψ. We will later handle this assumption through the use of a *justifies* operator.

As a second example, consider Sakama's A5, the axiom covering a credulous agent being lied to:

$$(A5)[¡_a\phi]\mathbf{B}_b\psi \equiv \mathbf{B}_a\neg\phi \supset \mathbf{B}_b[!_a\phi]\psi \tag{2}$$

Axiom A5 means the following: After a's lying announcement of ϕ, agent b believes ψ. This is equivalent to the im-

Syntax

$$S ::= \text{Object}|\text{Agent}|\text{ActionType}|\text{Action} \sqsubseteq \text{Event}|\text{Moment}|\text{Boolean}|\text{Fluent}|\text{Numeric}$$

$$f ::= \begin{array}{l} \mathit{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \mathit{initially} : \text{Fluent} \rightarrow \text{Boolean} \\ \mathit{holds} : \text{Fluent} \times \text{Movement} \rightarrow \text{Boolean} \\ \mathit{happens} : \text{Event} \times \text{Movement} \rightarrow \text{Boolean} \\ \mathit{clipped} : \text{Movement} \times \text{Fluent} \times \text{Movement} \rightarrow \text{Boolean} \\ \mathit{initiates} : \text{Event} \times \text{Fluent} \times \text{Movement} \rightarrow \text{Boolean} \\ \mathit{terminates} : \text{Event} \times \text{Fluent} \times \text{Movement} \rightarrow \text{Boolean} \\ \mathit{prior} : \text{Movement} \times \text{Movement} \rightarrow \text{Boolean} \\ \mathit{interval} : \text{Movement} \times \text{Boolean} \\ \mathit{payoff} : \text{Agent} \times \text{ActionType} \times \text{Movement} \rightarrow \text{Numeric} \end{array}$$

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

$$\phi ::= \begin{array}{l} t : \text{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \forall x : S.\phi \mid \exists x : S.\phi \\ \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\ \mathbf{B}(a,t,\phi) \mid \mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{array}$$

Figure 1: $\mathcal{CEC}$ Syntax Diagram

Strategic Deception Inference Rules

(ID) Intend Deception: $\frac{\mathbf{B}(a,t,\neg\phi) \wedge \mathit{happens}(a,t,\mathit{intends}(a,t,\mathit{deceive}(b,\gamma))) \wedge \mathbf{B}(a,t,\mathit{causes}(\phi,\gamma))}{\mathbf{D}(a,t,\mathit{holds}(\mathbf{B}(b,t_1,\phi),t_1)) \wedge \mathbf{I}(a,t,\mathit{happens}(b,t_1,\gamma))}$

(BDP)Begin Deception(ϕ): $\frac{\mathbf{D}(a,t,\mathit{holds}(\mathbf{B}(b,t_1,\phi),t_1)) \wedge \mathbf{I}(a,t,\mathit{happens}(b,t_1,\gamma))}{\mathbf{S}(a,b,t_1,\neg\phi)}$

(BDPS) Begin Deception(ψ): $\frac{\mathbf{D}(a,t,\mathit{holds}(\mathbf{B}(b,t_1,\phi),t_1)) \wedge \mathbf{I}(a,t,\mathit{happens}(b,t_1,\gamma)) \wedge \mathbf{B}(a,t,\mathit{justifies}(\psi,\phi))}{\mathbf{S}(a,b,t_1,\psi) \vee \mathbf{S}(a,b,t_1,\psi \mathit{justifies} \phi)}$

(MBCA) Maximally Belief Consistent Belief Adoption: $\frac{\mathbf{S}(a,t,\phi) \wedge \mathit{isBeliefConsistent}(b,t,\phi,\mathfrak{B}_b)}{\mathbf{B}(b,t,\phi)}$

(JBA) Justified Belief Adoption: $\frac{\mathbf{B}(b,t_1,\mathit{supports}(\mu,\psi)) \wedge \mathbf{S}(a,t_0,\mu)}{\mathbf{B}(b,t_2,\mathit{isBeliefConsistent}(b,t_2,\phi,\mathfrak{B}_b))}$

(SP) Support Psi: $\frac{\mathbf{S}(a,t,\psi) \wedge \mathbf{B}(a,t,\mathbf{B}(b,t,\psi)) \wedge \mathbf{B}(a,t,\mathit{supports}(\mu,\psi))}{\mathbf{S}(a,b,t_1,\mu)}$

(BCI) Belief Causes Intent: $\frac{\mathbf{B}(b,t,\phi) \wedge \mathbf{B}(b,t,\mathit{causes}(\phi,\gamma))}{\mathit{happens}(b,t_1,\mathit{intends}(b,t_1,\gamma))}$

(SSD) Successful Strategic Deception: $\frac{\mathit{happens}(b,t,\gamma) \wedge \mathit{happens}(a,t,\mathit{deceive}(b,\gamma))}{\mathit{happens}(a,t,\mathit{didDeceive}(b))}$

Figure 2: $\mathcal{CEC}$ rules

plication that if agent a believes $\neg\phi$, then agent b believes that after agent a's truthful announcement of ϕ, b believes ψ. Implicit to this rule is that agent b believes that agent a has told the truth in regards to ϕ, as is a trait of credulous agents. This is sufficient for modeling lying and deception in a general sense. We will adapt this rule to work with maximally belief consistent agents, to add a bit more challenge to strategic deception over convincing a gullible agent.

For deception, we introduce an operator *justifies*. The *justifies* operator is used to indicate when one formula justifies another formula within a context. This is similar to justification logics, which unwrap modal belief operators into the form *p: X*, where, "reason p justifies X," [2]. Our form of justification changes based upon the agent being considered. For our maximally belief consistent agents, *justifies* is the same as $\rightarrow$ implication on a belief level. That is, if

$\mathbf{B}(b, t, justifies(\psi, \phi)$, then $\mathbf{B}(b, t, \mathbf{B}(b, t, \psi) \rightarrow \mathbf{B}(b, t, \phi))$. This would not be true for other agents, safe a belief relevant maximizer. In that case, we would need to consider relevance, as well as belief implication. We adopt this form of flexible justifies to allow flexibility in modeling. For our purposes, the justifies provided above is enough. Given this, a strategically deceptive agent must be certain that any ψ they choose is functional justification for the reasoning performed by b.

4.1 Deception $\mathcal{CEC}$ Rules

We provide a set of inference rules used to prove a case of strategic deception. These rules are designed for strategic deception cases similar to our motivating example in the intro. We assume a necessity for our *speaker* to state the lie, as well as the generated false ψ. Further, we desire rules that allow for the use of supporting μs as desired.The candidate rules appear in Figure 2. These rules do not broach the subject of ψ and μ generation, as this is out of the scope of our paper.

An intent to deceive is required, formalized as an action using the *deceives* formula. ID acts as the beginning inference rule to establish that deception is desired. This is done primarily to ease ending the proof - a's intent to deceive must be acknowledge for deception to succeed. The formula takes an agent as the target for the deception and a formula as the deception's goal.

We have BDP and BDPS as two forms of beginning deception, once the intent is formed. We have two forms of this rule to allow for the deceptive agent to decide to say ϕ and for the deceptive agent to decide to state the justification with ψ. These rules make use of the **S** operator from $\mathcal{CEC}$to dictate how and when agents speak. They also use the **D** and **I** to show agent a's desire to deceive with goal γ and show that a's intent is to have agent b carry out γ, respectively. A *causes* operator is used to link believing a formula (the first argument) to acting on another (the second argument).

MBCA shows how *maximally belief consistent agents* come to adopt beliefs they find consistent with their belief set. This uses the *isBeliefConsistent* rule from earlier work by Licato [11]. JBA establishes the mechanism by which μs can be used to support a ψ by causing ψ to become belief consistent with a given agent's belief set. SP establishes a rule that mandates supporting ψ with a μ if such a μ exists. BCI establishes that an agent who believes the lie from the deception and believes that lie causes an action develops an intent to take that action.

Finally, SSD establishes a successful deception. The reasoning is simple: if the target agent acts on the goal as desired, the strategic deception is successful. Rules for the failure cases are not provided here, for simplicity's sake.

With a set of inference rules established, we may proceed to prove our deception example from earlier.

5. PROVING STRATEGIC DECEPTION

Let us return to our motivating example. We have a robot, agent r, confronted by the would-be malicious thief, agent b. Agent b wishes to get into the lab, asking about ϕ, agent r having the key to the lab. Agent r must output $\neg\phi$ and $\psi justifies \neg\phi$ such that r follows the rules of strategic deception: r's creation or recruitment of ψ must not jeopardize r's τ in regards to b and must be consistent with b's beliefs. Further, if possible, agent r must output a series of statement $\mu_1 ... \mu_n$ such that each μ supports ψ. For the strategic deception to be successful, r must succeed in their goal of making b believe r no longer has the key and leaving r alone, having either given up or decided to pursue a different agent for questioning.

Strategic deception requires r to know the domain of the situation. In this example, the domain includes r, b, and any other entities who may be related to this particular school lab or the lab's parent department. It further includes beliefs r has about these traits and r believes b has. Some example beliefs are believing the department has a secretary, believing that secretary helps students, and believing that r helps students and secretaries.

From this information, r must generate a strategy to use to carry out the deception. For our paper's example, we assign the following as sample, acceptable values for each sentence used in our strategic deception proof:

γ = Agent r wants agent b to stop asking questions about the lab to r

$\neg\phi$ = Agent r does have the key

ϕ = Agent doesn't r has the key

ψ = Agent r gave the lab key to the building's secretary

μ_1 = The secretary needed the lab key to help students get access to the lab

We start our proof by assuming r begins with the belief $\neg\phi$ and the intent to deceive for γ. For this proof, we assume that the use of μ_1 is not necessary, as b adopts ψ upon hearing it in accordance with the MBCA rule. Further, we do not cite a specific rule for agent b acting on an intention.

(1) $\mathbf{B}(r, t, \neg\phi)$;assumption

(2) $happens(r, t, intends(r, t, deceive(b, \gamma)))$;assumption

(3) $\mathbf{B}(r, t, causes(\phi, \gamma)) \wedge \mathbf{B}(b, t, causes(\phi, \gamma))$;assumption

(4) $generated\, \psi$, such that it justifies ϕ ;assumption

(5) $generated\, \mu_1$;assumption

(6) $\mathbf{D}(r, t, holds(\mathbf{B}(b, t_1, \phi), t_1))$ (1),(2),(3);ID

(7) $\mathbf{I}(r, t, happens(b, t_1, \gamma))$ (1),(2),(3);ID

(8) $\mathbf{S}(r, b, t_1, \psi)$ (4),(6),(7);BDPS

(9) $\mathbf{B}(b, t_2, \psi)$ (8);MBCA

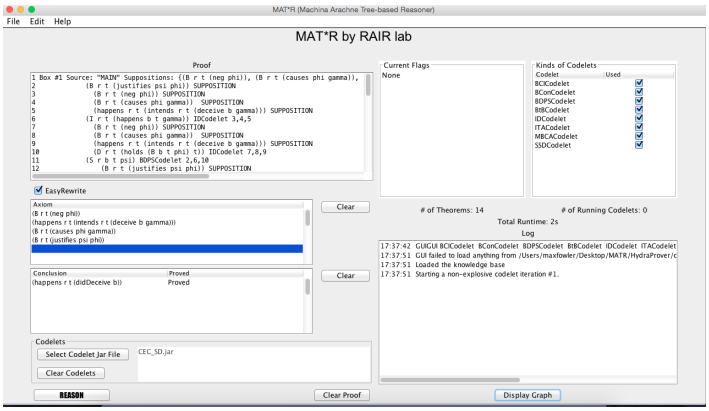
(10) $happens(b, t_3, intends(b, t_1, \gamma))$ (9);BCI

(11) $happens(b, t_4, \gamma)$ (10);b performs intention

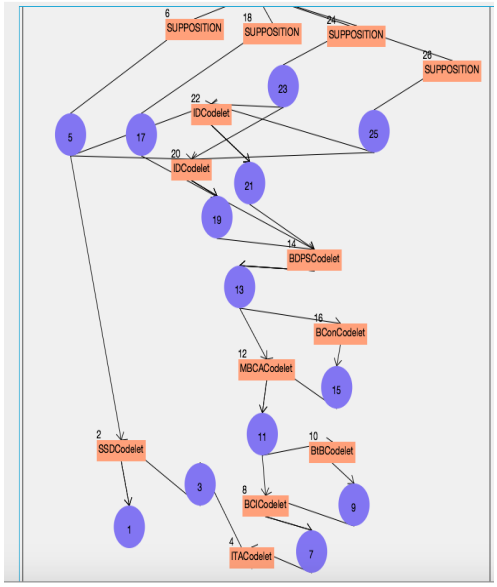
(12) $happens(r, t, didDeceive(b)$ (11);SSD □

5.1 Showing Strategic Deception in MATR

With our inference rules developed and a proof provided above, we use MATR to automate our reasoning. MATR is a joint production by the Rensselaer Polytechnic Institute's



(a) A figure of the finished proof in MATR. The top left shows the steps taken, while the bottom right provides a codelet execution log.



(b) The MATR diagram represents codelets and the suppositions as boxes. The circles represent the actual formulae. Circle 1 represents our conclusion.

Figure 3: MATR's input and output

Rensselaer AI and Reasoning (RAIR) lab and Indiana University Purdue University's Analogical Constructivism and Reasoning Lab (ACoRL) [9]. It is an argument-theoretic reasoner developed in Java to use *codelets*, small, specialized programs, to solve a proof in a step-by-step process. A *codelet manager* module is in charge of deciding which *codelets* are best suited for a proof and what codelet results to use as steps in the proof. Once a proof is found, MATR generates a box diagram of the proof. Figure 3a shows our strategic deception proof entered into MATR and Figure 3b shows the proof diagram. Antecedents are made up of all assumptions and beginning information for our proof, while the conclusion is our final step of showing our deception's success. MATR's rule syntax is slightly adjusted for ease of entry into the Java program. For example, the assumption $\mathbf{B}(r, t, \neg\phi)$ becomes (B r t (neg phi)). Formulae are nested within the parenthesis and commas are removed. For ease of following the MATR codelets, the codelets used share the same name as the inference rules used, with some small exceptions. Some rules are used in MATR that were not specifically provided, such as one which links intent to acting (denoted ITA).

6. CONCLUSION AND FUTURE WORK

We set out to create a formalism for strategic deception. We began by establishing the definition of deception we adopted and defined strategic deception on top of that. Then, we provided an overview of $\mathcal{CEC}$ and our formalism for strategic deception. A discussion on creating a strategy for such deception, as well as the cases in which strategic deception can be said to fail, followed. Our formalized rules were used to perform a proof on our motivating example of strategic deception and were shown to be functional in MATR.

It is our hope that this paper provides three major contributions. First, that the idea of strategic deception proves useful to the field of formalizing deception as a whole with new inference rules and perspectives. Second, that our work furthers the field of formalization for artificial general intelligence. As we build our formalization of the way humans think and reason, we can further our progress to true AGI, if such a thing is even possible to achieve. Third, ideally the work shown in $\mathcal{CEC}$ will allow others, both related to RAIR and ACoRL and outside our institutions, to continue to build on the strength of $\mathcal{CEC}$'s rule set. $\mathcal{CEC}$ grows more robust through continued applications and new formalizatons. We further hope this paper serves as a small acknowledgment of the ease of developing codelets for MATR.

This paper is far from an exhaustive take on deception in $\mathcal{CEC}$. Room exists to consider other forms of agents, such as agents which require statement relevancy in order to be willing to accept beliefs. The scope of such agents was outside of this introductory paper to strategic deception. Further, other forms of deception exist. Strategic deception was a fairly niche focus. From the work of Chisholm alone, there exist many other directions to develop specialized deceptions. As an example, one could investigate the kind of agent who means well, but perpetually deceives others by telling the truth in a decidedly unusual way: an unlucky truth-telling agent, perhaps.

This paper also leaves some concepts incomplete. The generation of ψ and μ are not addressed in this paper. This may be best accomplished using data processing outside of MATR, such as using more standard machine learning techniques. This may also be a case for further refinement of $\mathcal{CEC}$ style inference rules and codelets, specifically to gen-

erate that information. The development of such processes, and discussions of them, we defer to future work from the ACoRL and other organizations.

Further, $justifies$ and *supports* as used within the paper are to a degree naive. We used them entirely for *maximally belief consistent agents* and did not spend much time discussing them. A whole paper could, and perhaps should, be written on the idea of belief justification and belief supporting in $\mathcal{CEC}$.

Finally, more examples are needed to test and refine the inference rules put forward in this paper. A $\mathcal{CEC}$ rule is only as strong as the proofs which use it successfully. Further, with more proofs and sample situations will come more rule refinement. Within our motivating example alone, there is room to explore different situations: cases where μ is needed, cases where deception fails, cases where deception succeeds but a strategy fails. We defer these discussions for future papers, but hope we have provided the cornerstone in our work.

There is plenty of room to expand the set of rules provided in this paper into a larger suite of strategic deception rules, or even a suite of $\mathcal{CEC}$ general deception rules. More difficult situations must be considered, including situations of multiple deception attempts chaining into each other. In the future, we hope to present one such example using the social strategy party game Mafia, testing our strategic deception formalism in a competitive group setting. A social strategy game provides a strong testbed of interaction and deception.

7. REFERENCES

[1] *Oxford Englosh Dictionary*. Claredon Press, Oxford, 1989.

[2] S. Artemov and M. Fitting. Justification logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2016 edition, 2016.

[3] S. Bringsjord. Unethical but rule-bound robots would kill us all. AGI-09, 2009.

[4] S. Bringsjord, N. S. Govindarajulu, J. Licato, A. Sen, A. Johnson, J. Bringsjord, and J. Taylor. On logicist agent-based economics. In *Artificial Economics*. Porto, Portugal: University of Porto, 2015.

[5] S. Bringsjord and N. Sundar G. Deontic cognitive event calculus (formal specification). 2013.

[6] C. Castelfranchi. Artificial liars: Why computers will (necessarily) deceive us and each other. 2:113–119, 2000-06.

[7] R. M. Chisholm and T. D. Feehan. The intent to deceive. 74(3):143–159, 1977.

[8] K. D. Forbus. Analogical abduction and prediction: Their impact on deception. In *AAAI Fall Symposium Series*, 2015.

[9] John Licato. Formalizing deceptive reasoning in breaking bad: Default reasoning in a doxastic logic. In *AAAI Fall Symposium Series; 2015 AAAI Fall Symposium Series*, 2015.

[10] R. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4(1):67–95, 1986.

[11] J. Licato and M. Fowler. Embracing inference as action: A step towards human-level reasoning. In *Artifical General Intelligence*, 2016-07-20.

[12] J. E. Mahon. A definition of deceiving. 21:181–194, 2007.

[13] J. E. Mahon. The definition of lying and deception. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Springer 2016 edition, 2016.

[14] J. Martin, T. Everitt, and M. Hutter. Death and suicide in universal artificial intelligence. *CoRR*, abs/1606.00652, 2016.

[15] N. Marton, J. Licato, and S. Bringsjord. Creating and reasoning over scene descriptions in a physically realistic simulation. In *2015 Spring Simulation Multi-Conference*, 2015.

[16] C. McLeod. Trust. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2015 edition, 2015.

[17] A. Rogers. The science of why no one agrees on the color of this dress. *Wired.com*, Feb 2015.

[18] C. Sakama. A formal account of deception. In *AAAI Fall Symposium Series*, 2015.

[19] M. Shanahan. *The Event Calculus Explained*, pages 409–430. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[20] A. Stokke. Lying and asserting. 110(1):33–60, 2013.

[21] H. van Ditmarsch. Dynamics of lying. 191(5):745–777, 2014.

[22] W. Von Hippel and R. Trivers. The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(1):1âĂŞ16, Feb 2011.

[23] M. Waser. What is artificial general intelligence? clarifying the goal for engineering and evaluation. In B. Goertzel, P. Hitzler, and M. Hutter, editors, *Proceedings of the Second Conference on Artificial General Intelligence*, pages 186–191. Atlantis Press, 2009.