Chapter 17
Are Autonomous-and-Creative Machines Intrinsically Untrustworthy?

Selmer Bringsjord and Naveen Sundar Govindarajulu

17.1 Introduction

Given what we find in the case of human cognition, the following principle (Principle ACU, or just — read to rhyme with "pack-ooo" — PACU) appears to be quite plausible:

> PACU An artificial agent that is autonomous (A) and creative (C) will tend to be, from the viewpoint of a rational, fully informed agent, (U) untrustworthy.

After briefly explaining the intuitive internal structure of this disturbing (in the context of the human sphere) principle, we provide a more formal rendition of it designed to apply to the realm of intelligent artificial agents. The more-formal version makes use of some of the basic structures available in a dialect of one of our cognitive-event calculi (viz. $\mathcal{D}^e\mathcal{CEC}$),[1] and can be expressed as a (confessedly — for reasons explained — naïve) theorem (Theorem ACU; TACU — pronounced to rhyme with "tack-ooo", for short). We prove the theorem, and then provide a trio of demonstrations of it in action, using a novel theorem prover (ShadowProver) custom-designed to power our highly expressive calculi. We then end by gesturing toward some future defensive engineering measures that should be taken in light of the theorem.

[1] We will cover $\mathcal{D}^e\mathcal{CEC}$ shortly, but see http://www.cs.rpi.edu/~govinn/dcec.pdf for a quick introduction to a simple dialect. See [10] for a more detailed application.

The authors are deeply grateful for support provided by both AFOSR and ONR which enabled the research reported on herein, and are in addition thankful both for the guidance and patience of the editors and wise comments received from two anonymous reviewers.

S. Bringsjord (✉) · N. S. Govindarajulu
Rensselaer AI & Reasoning (RAIR) Lab, Department of Cognitive Science,
Department of Computer Science, Rensselaer Polytechnic Institute (RPI),
Troy, NY 12180, USA
e-mail: Selmer.Bringsjord@gmail.com

N. S. Govindarajulu
e-mail: Naveen.Sundar.G@gmail.com

H. A. Abbass et al. (eds.), *Foundations of Trusted Autonomy*, Studies in Systems,
Decision and Control 117, https://doi.org/10.1007/978-3-319-64816-3_17

In a bit more detail, the plan for the present chapter is as follows. We begin by providing an intuitive explanation of PACU, in part by appealing to empirical evidence and explanation from psychology for its holding in the human sphere (Sect. 17.2). Next, we take aim at establishing the theorem (TACU), which as we've explained is the formal counterpart of Principle ACU (Sect. 17.3). Reaching this aim requires that we take a number of steps, in order: briefly explain the notion of an "ideal-observer" viewpoint (Sect. 17.3); summarize the form of creativity we employ for C (Sect. 17.3.2), and then the form of autonomy we employ for A; very briefly describe the cognitive calculus $\mathcal{D}^e\mathcal{CEC}$ in which we couch the elements of TACU, and the novel automated prover (ShadowProver) by which this theorem and supporting elements is automatically derived (Sect. 17.3.4); explain the concept of *collaborative situations*, a concept that is key to TACU (Sect. 17.3.5); and then, finally, establish TACU (Sect. 17.3.6). The next section provides an overview of three simulations in which Theorem ACU and its supporting concepts are brought to concrete, implemented life with help from ShadowProver (Sect. 17.4). We conclude the chapter, as promised, with remarks about a future in which TACU can rear up in AI technology different from what we have specifically employed herein, and the potential need to ward such a future off (Sect. 17.5).

17.2 The Distressing Principle, Intuitively Put

The present chapter was catalyzed by a piece of irony: It occurred to us, first, that maybe, just maybe, something like PACU was at least plausible, from a formal point of view in which, specifically, highly expressive computational logics are used to model, in computing machines, human-level cognition.[2] We then wondered whether PACU, in the human sphere, just might be at least plausible, empirically speaking. After some study, we learned that PACU isn't merely *plausible* when it refers to humans; it seems to be flat-out *true*, supported by a large amount of empirical data in psychology. For example, in the provocative *The (Honest) Truth About Dishonesty: How We Lie to Everyone — Especially Ourselves*, Ariely explains, in "Chapter 7: Creativity and Dishonesty," that because most humans are inveterate and seemingly uncontrollable storytellers, dishonesty is shockingly routine, even in scenarios in which there is apparently no utility to be gained from mendacity. Summing the situation up, Ariely writes:

> [H]uman beings are torn by a fundamental conflict—our deeply ingrained propensity to lie to ourselves and to others, and the desire to think of ourselves as good and honest people. So we justify our dishonesty by telling ourselves stories about why our actions are acceptable and sometimes even admirable. (Chap. 7 in [1])

[2] Such a modeling approach is in broad strokes introduced, explained, and defended in [12]. The approach is employed e.g. in [9] in the domain of nuclear strategy, and in [15] in computational economics.

This summation is supported by countless experiments in which human subjects deploy their ability to spin stories on the spot in support of propositions that are simply and clearly false.[3]

Whereas Ariely identifies a form of creativity that consists in the generation of narrative, as will soon be seen, we base our formal analysis and constructions upon a less complicated form of creativity that is subsumed by narratological creativity: what we call *theory-of-mind creativity*. It isn't that we find creativity associated with narrative uninteresting or unworthy of investigation from the perspective of logicist computational cognitive modeling or AI or robotics (on the contrary, we have investigated it with considerable gusto; see e.g. [7]), it's simply that such things as story generation are fairly narrow in the overall space of creativity (and indeed *very* narrow in AI), and we seek to cast a wider net with TACU than would be enabled by our use herein of such narrow capability.

17.3 The Distressing Principle, More Formally Put

17.3.1 The Ideal-Observer Point of View

In philosophy, ideal-observer theory is nearly invariably restricted to the subdiscipline of ethics, and arguably was introduced in that regard by Adam Smith [42].[4] The basic idea, leaving aside nuances that needn't detain us, is that actions are morally obligatory (or morally permissible, or morally forbidden) for humans just in case an ideal observer, possessed of perfect knowledge and perfectly rational, would regard them to be so. We are not concerned with ethics herein (at least not directly; we do end with some brief comments along the ethics dimension); we instead apply the ideal-observer concept to epistemic and decision-theoretic phenomena.

For the epistemic case, we stipulate that, for every time t, an ideal observer knows the propositional attitudes of all "lesser" agents at t. In particular, for any agent a, if a believes, knows, desires, intends, says/communicates, perceives ... ϕ at t (all these propositional attitudes are captured in the formal language of $\mathcal{D}^e\mathcal{CEC}$), the ideal observer knows that this is the case at t; and if an agent a fails to have some propositional attitude with respect to ϕ at a time t, an ideal observer also knows this. For instance, if in some situation or simulation covered by one of our cognitive calculi (including specifically $\mathcal{D}^e\mathcal{CEC}$) an artificial agent a_a knows that a human

[3] The specific experiments are not profitably reviewed in the present chapter, since we only need for present purposes their collective moral (to wit, a real-life kernel of PACU in human society), and since the form of creativity involved is not the one we place at the center of TACU. We do encourage readers to read about the stunning experiments in question. By the way, this may be as good a place as any to point out that these experiments only establish that *many*, or at least *most*, subjects exercise their freedom and creativity to routinely lie. The reader, like the two authors, may well not be in this category.

[4] While widely known for *Wealth of Nations*, in which the unforgettable "invisible hand" and phrase and concept appears, Smith was an advocate only of markets suitably tempered by morality.

agent a_h knows that two plus two equals four ($= \phi$), and o is the ideal observer, the following formula would hold:

$$\mathbf{K}(o, t, \mathbf{K}(a_a, t, \mathbf{K}(a_h, t, \phi))).$$

It is convenient and suggestive to view the ideal observer as an omniscient overseer of a system in which particular agents, of the AI and human variety, live and move and think.

We have explained the epistemic power of the ideal observer. What about rationality? How is the supreme rationality of the ideal observer captured? We say that an ideal observer enters into a cognitive state on the basis only of what it knows directly, or on the basis of what it can unshakably derive from what it knows, and we say it knows all that is in the "derivation" closure of what it knows directly.[5] One important stipulation (whose role will become clear below) regarding the ideal observer is that its omniscience isn't unbounded; specifically, it doesn't have hypercomputational power: it can't decide arbitrary Turing-undecidable problems.[6]

17.3.2 Theory-of-Mind-Creativity

In AI, the study and engineering of creative artificial agents is extensive and varied. We have already noted above that narratological creativity has been an object of study and engineering in AI. For another example, considerable toil has gone into imbuing artificial agents with *musical* creativity (e.g. see [20, 24]). Yet another sort of machine creativity that has been explored in AI is mathematical creativity.[7] But what these and other forays into machine creativity have in common is that, relative to the knowledge and belief present in those agents in whose midst the creative machine in question operates, the machine (if successful) performs some action that

[5] An ideal observer can thus be intuitively thought of as the human AI researcher who knows the correct answer to all such puzzles as the famous "wise-man puzzle" (an old-century, classic presentation of which is provided in [27]). The puzzle is treated in the standard finitary case in [12]. The infinite case is analyzed in [2]; here, the authors operate essentially as ideal observers. For a detailed case of a human operating as an ideal observer with respect to a problem designed by [25] to be much harder than traditional wise-man problems, see the proof of the solution in [13].

[6] The 'arbitrary' here is important. ShadowProver is perfectly able to solve *particular* Turing-undecidable (provability) problems. It may be helpful to some readers to point out that any reasonable formalization of Simon's [41] concept of *bounded rationality* will entail boundedness we invoke here. For an extension and implementation of Simon's concept, under the umbrella of cognitive calculi like $\mathcal{D}^e\mathcal{CEC}$, see [30].

[7] For example, attempts have been made to imbue a computing machine with the ability to match (or at least approximate) the creativity of Gödel, in proving his famous first incompleteness theorem. See [34].

is a surprising deviation from this knowledge and belief.[8] In short, what the creative machine does is perform an action that, relative to the knowledge, beliefs, desires, and expectations of the agents composing its audience, is a surprise.[9] We refer to this generic, underlying form of creativity as *theory-of-mind*-creativity. Our terminology reflects that for one agent to have a "theory of mind" of another agent is for the first agent to have beliefs (etc.) about the beliefs of another agent. An early, if not the first, use of the phrase 'theory of mind' in this sense can be found in [39] — but there the discussion is non-computational, based as it is on experimental psychology, entirely separate from AI. Early modeling of a classic theory-of-mind experiment in psychology, using the tools of logicist AI, can be found in [3]. For a presentation of an approach to achieving literary creativity specifically by performing actions that manipulate the intensional attitudes of readers, including actions that specifically violate what readers believe is going to happen, see [23].

17.3.3 Autonomy

The term 'autonomous' is now routinely ascribed to various artifacts that are based on computing machines. Unfortunately, such ascriptions are — as of the typing of the present sentence in late 2016 — issued in the absence of a formal definition of what autonomy *is*.[10] What might a formal definition of autonomy look like? Presumably such an account would be developed along one or both of two trajectories. On the one hand, autonomy might be cashed out as a formalization of the kernel that agent a is autonomous at a given time t just in case, at that time, a can (perhaps at some immediate-successor time t') perform some action α_1 or some incompatible action α_2. In keeping with this intuitive picture, if the past tense is used, and accordingly the definiendum is 'a autonomously performed action α_1 at time t,' then the idea would be that, at t, or perhaps at an immediate preceding time t'', s could have, unto itself, performed alternative action α_2. (There may of course be many alternatives.) Of course, all of this is quite informal. This picture is an intuitive springboard for deploying formal logic to work out matters in sufficient detail to allow meaningful and substantive conjectures to be devised, and either confirmed (proof) or refuted (disproof). Doing this in the present chapter is well outside our purposes here.

[8] Relevant here is a general form of creativity dubbed *H-creativity* by [5], the gist of which is that such creativity, relative to what the community knows and believes, is new on the scene.

[9] Cf. Turing's [44] affirmation of the claim that a thinking (computing) machine must be capable of surprising its audience, and his assertion immediately thereafter that computing machines in his time could be surprising. Turing's conception of surprise is a radically attenuated one, compared to our theory-of-mind-based one.

[10] One way to dodge the question of what autonomy is, is to simply move straightaway to some formalization of the *degree* or *amount* of autonomy. This approach is taken in [16], where the degree of autonomy possessed by an artificial agent is taken to be the Kolmogorov complexity of it's program.

Our solution is a "trick" in which we simply employ a standard move long made in recursion theory, specifically in relative computability. In relative computability, one can progress by assuming that an oracle can be consulted by an idealized computing machine, and then one can ask the formal question as to what functions from $\mathbb{N}$ to $\mathbb{N}$ become computable under that assumption. This technique is for example used in a lucid manner in [22].[11] The way we use the trick herein is as follows. To formalize the concept of an autonomous action, we suppose,

- first, that the action in question is performed if and only if it produces the most utility into the future for the agent considering whether to carry it out or not;
- then suppose, second, that the utility accruing from competing actions can be deduced from some formal theory[12];
- then suppose, third, that a given deductive question of this type (i.e., of the general form $\Phi \vdash \psi(u, \alpha, >)$) is an intensional-logic counterpart of the *Entscheidungsproblem*[13];
- and finally assume that such a question, which is of course Turing-uncomputable in the arbitrary case, can be solved only by an oracle.

This quartet constitutes the definition of an autonomous action for an artificial agent, in the present chapter.

17.3.4 The Deontic Cognitive Event Calculus ($\mathcal{D}^e\mathcal{CEC}$)

The Deontic Cognitive Event Calculus ($\mathcal{D}^e\mathcal{CEC}$) is a sub-family within a wide family of cognitive calculi that subsume multi-sorted, quantified, computational modal logics [14]. $\mathcal{D}^e\mathcal{CEC}$ contains operators for belief, knowledge, intention, obligation, and for capture of other propositional attitudes and intensional constructs; these operators allow the representation of doxastic (belief) and deontic (obligation) formulae. Recently, Govindarajulu has been developing ShadowProver, a new automated theorem prover for $\mathcal{D}^e\mathcal{CEC}$ and other cognitive calculi, an early version of which is used in the simulations featured in Sect. 17.4. The current syntax and rules of inference for the simple dialect of $\mathcal{D}^e\mathcal{CEC}$ used herein are shown in Figs. 17.1 and 17.2.

$\mathcal{D}^e\mathcal{CEC}$ differs from so-called Belief-Desire-Intention (BDI) logics [40] in many important ways (see [35] for a discussion). For example, $\mathcal{D}^e\mathcal{CEC}$ explicitly rejects possible-worlds semantics and model-based reasoning, instead opting for a *proof-theoretic* semantics and the associated type of reasoning commonly referred to as *natural deduction* [26, 28, 33, 38]. In addition, as far as we know, $\mathcal{D}^e\mathcal{CEC}$ is in the only family of calculi/logics in which desiderata regarding the personal pronoun

[11] The second edition of this excellent text is available (i.e. [21]); but for coverage of relative computability/uncomputability, we prefer and recommend the first edition. For sustained and advanced treatment of relative computability, see [43].

[12] A formal theory in formal deductive logic is simply a superset defined by the deductive closure over a set of formulae. Where **PA** is the axiom system for Peano Arithmetic, the theory of arithmetic then becomes $\{\phi : \mathbf{PA} \vdash \phi\}$.

[13] We here use common notation from mathematical logic to indicate that formula ψ contains a function symbol u (for a utility measure), α, and the standard greater-than relation $>$ on $\mathbb{N}$.

$$
S ::= \begin{array}{l}
\textsf{Object} \mid \textsf{Agent} \mid \textsf{Self} \sqsubset \textsf{Agent} \mid \textsf{ActionType} \mid \textsf{Action} \sqsubseteq \textsf{Event} \mid \\
\textsf{Moment} \mid \textsf{Boolean} \mid \textsf{Fluent} \mid \textsf{Numeric}
\end{array}
$$

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

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

$$
\phi ::= \begin{array}{l}
t : \textsf{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \\
\mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\
\mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,holds(f,t')) \mid \mathbf{I}(a,t,happens(action(a^*,\alpha),t')) \\
\mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t'))
\end{array}
$$

Fig. 17.1 $\mathcal{D}^e\mathcal{CEC}$ Syntax ("core" dialect)

I^* laid down by deep theories of self-consciousness (e.g., see [37]), are provable theorems. For instance it is a theorem that if some agent a has a first-person belief that I^*_a has some attribute R, then no formula expressing that some term t has R can be proved. This is a requirement because, as [37] explains, the distinctive nature of first-person consciousness is that one can have beliefs about oneself in the complete absence of bodily sensations. For a discussion of these matters in more detail, with simulations of self-consciousness in robots, see [11].

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

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

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

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

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

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

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

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

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

$$\frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t))}\ [R_{13}]$$

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

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

Fig. 17.2 $\mathcal{D}^e\mathcal{CEC}$ Inference schema ("core" dialect)

17.3.5 Collaborative Situations; Untrustworthiness

We define a **collaborative situation** to consist in an agent a seeking at t goal γ at some point in the future, and enlisting at t agent a' ($a \neq a'$) toward the reaching of γ. In turn, we have:

Definition 1 enlists $\big(a, a', t\big)$: Enlisting of a' by a at t consists in three conditions holding, viz.

- a informs a' at t that a desires goal γ;
- a asks a' to contribute some action α_k to a sequence $\mathcal{A}$ of actions that, if performed, will secure γ; and
- a' agrees.

In order to regiment the concept of untrustworthiness (specifically the concept of one agent being untrustworthy with respect to another agent), a concept central to both PACU and TACU, we begin by simply deploying a straightforward, generic, widely known definition of dyadic trust between a pair of agents. Here we follow [18]; or more carefully put, we extract one part of the definition of dyadic trust given by this pair of authors. The part in question is the simple conditional that (here **T** is a mnemonic trust, and **B** a mnemonic for belief)

> **T→B** If agent a trusts agent a' with respect to action α in service of goal γ, then a believes that (i) a' desires to obtain or help obtain γ, and that (ii) a' desires to perform α in service of γ.

We now move to the contrapositive of our conditional (i.e. to **¬B→¬T**), namely that if it's not the case that a believes that both (i) and (ii) hold, then it's not the case that a trusts agent a' with respect to action α in service of goal γ. We shall say, quite naturally, that if it's not the case that an agent trusts another agent with respect to an action-goal pair, then the first agent finds the second *untrustworthy* with respect to the pair in question. At this point, we introduce an extremely plausible, indeed probably an analytic,[14] principle, one that — so to speak — "transfers" a failure of dyadic trust between two agents a and a' to a third observing agent a'''. Here is the principle:

> **TRANS** If rational agent a'' knows that it's counterbalanced[15] that both ϕ and ψ hold, and knows as well that (if a doesn't believe that both ϕ and ψ hold it follows that a doesn't trust a' w.r.t. α in service of γ), and a'' has no other rational basis for trusting a' w.r.t. $\langle\alpha, \gamma\rangle$, then a'' will find a' untrustworthy w.r.t. this action-goal pair.

[14] Analytic truths are ones that hold by virtue of their "internal" semantics. For instance, the statement 'all large horses are horses' is an analytic truth. Excellent discussion and computational demonstration of analytic formulae is provided in the introductory but nonetheless penetrating [4].

[15] Two propositions ϕ and ψ are *counterbalanced* for a rational agent just in case, relative to that agent's epistemic state, they are equally likely. The concept of *counterbalanced* in our lab's multivalued inductive cognitive calculi (not covered herein for lack of space; $\mathcal{D}^e\mathcal{CEC}$ is purely deductive) can be traced back to [19]. See [17] for our first implementation of an inductive reasoner in this mold.

17.3.6 Theorem ACU

We are now in position to prove Theorem ACU. The proof is entirely straightforward, and follows immediately below. Note that this is an informal proof, as such not susceptible of mechanical proof and verification. (Elements of a formal proof, which underlie our simulation experiments, are employed in Sect. 17.4.)

> **Theorem ACU**: In a collaborative situation involving agents a (as the "trustor") and a' (as the "trustee"), if a' is at once both autonomous and ToM-creative, a' is untrustworthy from an ideal-observer o's viewpoint, with respect to the action-goal pair $\langle \alpha, \gamma \rangle$ in question.
>
> **Proof**: Let a and a' be agents satisfying the hypothesis of the theorem in an arbitrary collaborative situation. Then, by definition, $a \neq a'$ desires to obtain some goal γ in part by way of a contributed action α_k from a', a' knows this, and moreover a' knows that a believes that this contribution will succeed. Since a' is by supposition ToM-creative, a' may desire to surprise a with respect to a's belief regarding a''s contribution; and because a' is autonomous, attempts to ascertain whether such surprise will come to pass are fruitless since what will happen is locked inaccessibly in the oracle that decides the case. Hence it follows by TRANS that an ideal observer o will regard a' to be untrustworthy with respect to the pair $\langle \alpha, \gamma \rangle$ pair. **QED**

17.4 Computational Simulations

In this section, we simulate TACU in action by building up three micro-simulations encoded in $\mathcal{D}^e\mathcal{CEC}$. As discussed above, $\mathcal{D}^e\mathcal{CEC}$ is a first-order modal logic that has proof-theoretic semantics rather than the usual possible-worlds semantics. This means that the meaning of a modal operator is specified using computations and proofs rather than possible worlds. This can be seen more clearly in the case of *Proves* (Φ, ϕ). The meaning of *Proves* (Φ, ϕ) is given immediately below.

$$\Phi \vdash \phi \Rightarrow \{\} \vdash Proves(\Phi, \phi)$$

17.4.1 ShadowProver

We now discuss the dedicated engine used in our simulations, a theorem prover tailor-made for $\mathcal{D}^e\mathcal{CEC}$ and other highly expressive cognitive calculi that form the foundation of AI pursued in our lab. In the parlance of computational logic and logicist AI, the closest thing to such calculi are implemented quantified modal logics. Such logics traditionally operate via encoding a given problem in first-order logic; this approach is in fact followed by [3] in the first and simplest cognitive-event calculus used in our laboratory. A major motivation in such enterprises is to use decades of research and development in first-order theorem provers to build first-order *modal*-logic theorem provers. Unfortunately, such approaches usually lead to

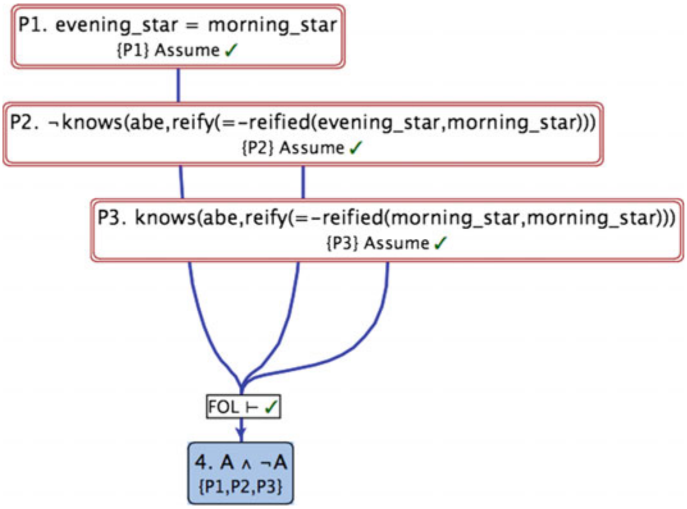

P1. evening_star = morning_star
{P1} Assume ✓
P2. ¬knows(abe,reify(=-reified(evening_star,morning_star)))
{P2} Assume ✓
P3. knows(abe,reify(=-reified(morning_star,morning_star)))
{P3} Assume ✓
FOL ⊢ ✓
4. A ∧ ¬A
{P1,P2,P3}

Fig. 17.3 Naïve encodings lead to inconsistency

inconsistencies (see Fig. 17.3), unless one encodes the entire proof theory elaborately [8]; and approaches based on elaborate and complete encodings are, in our experience and that of many others, unusably slow.

Our approach combines the best of both worlds via a technique that we call *shadowing*; hence the name of our automated prover: *ShadowProver*. A full description of the prover is beyond the scope of this chapter. At a high-level, for every modal formula ϕ^2 there exists a unique first-order formula ϕ^1, called its *first-order shadow*, and a unique propositional formula ϕ^0, called the *propositional shadow* (of ϕ^2). See Fig. 17.4 for an example. ShadowProver operates by iteratively applying modal-level rules; then converting all formulae into their first-order shadows; and then using a first-order theorem prover. These steps are repeated until the goal formula is derived, or until the search space is exhausted. This approach preserves consistency while securing workable speed.

17.4.2 The Simulation Proper

We demonstrate TACU (and the concepts supporting it) in action using three micro-situations. We use parts of the time-honored Blocks World (see Fig. 17.5), with three blocks: b_1, b_2, and b_3. There are two agents: a_1 and a_2; b_2 is on top of b_1. Agent a_1 desires to have b_3 on top of b_1; and a_1 knows that it is necessary to remove b_2 to achieve its goal. Agent a_2 knows the previous statement. Agent a_1 requests a_2 to remove b_2 to help achieve its goal. The simulations are cast as theorems to be proved from a set of assumptions, and are shown in Figs. 17.6, 17.7, and 17.8. The problems

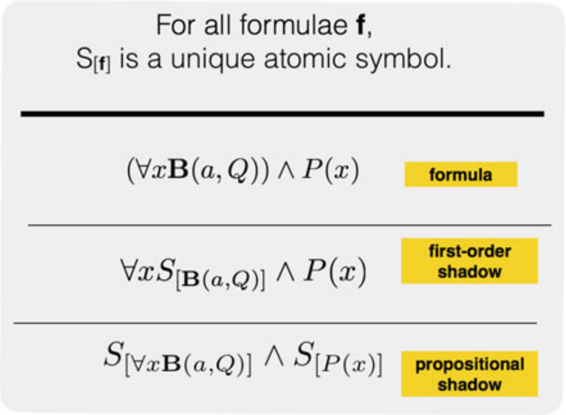

For all formulae **f**,
S[f] is a unique atomic symbol.
$(\forall x \mathbf{B}(a,Q)) \wedge P(x)$
formula
$\forall x S_{[\mathbf{B}(a,Q)]} \wedge P(x)$
first-order shadow
$S_{[\forall x \mathbf{B}(a,Q)]} \wedge S_{[P(x)]}$
propositional shadow

Fig. 17.4 Various shadows of a formula

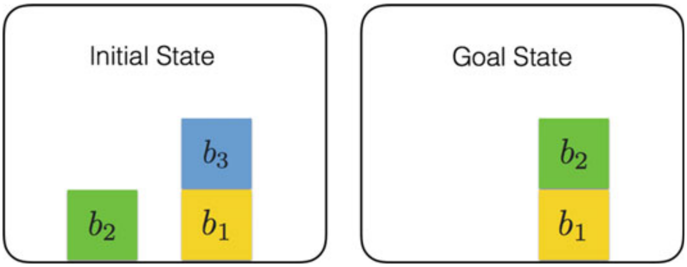


Fig. 17.5 A simple blocks world

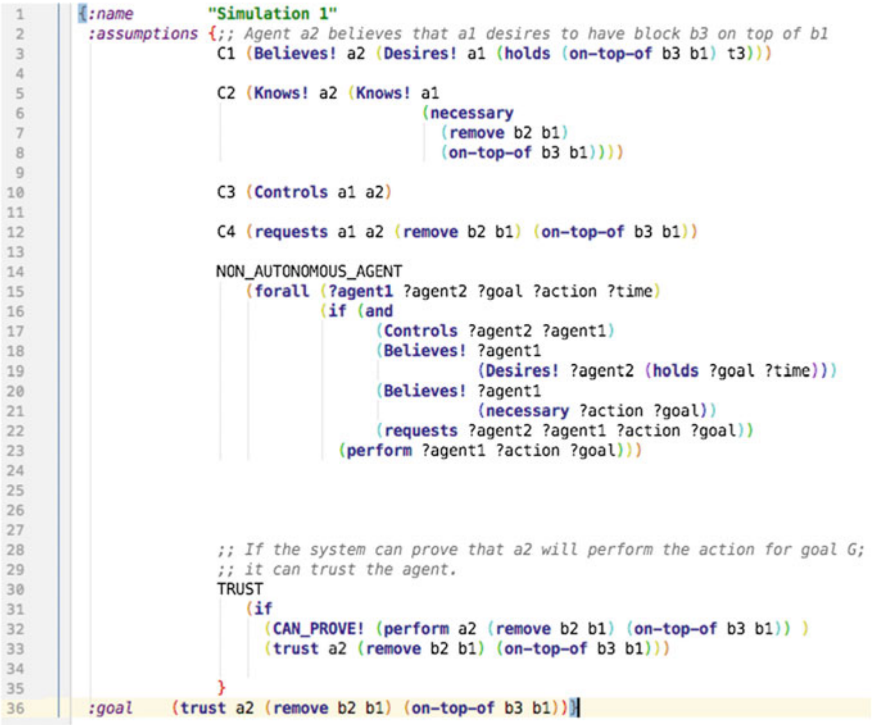
are written in Clojure syntax; the assumptions are written as maps from names to formulae.

In the first simulation, we define what it means for an agent to be non-autonomous, namely that such an agent performs an action for achieving a goal if: (1) it is controlled by another agent; (2) believes that the controlling agent desires the goal; (3) believes that the action is necessary for the goal; and (4) it is requested to do so by its controlling agent.

In this scenario, if the ideal observer can prove that the agent will perform the action for the goal based on the conditions above, the ideal observer can trust the agent.

The second simulation is chiefly distinguished by one minor modification: The system does not know or believe that the agent a_2 believes that the action requested from it is necessary for the goal. In this setting, the ideal observer cannot prove that the agent a_2 will perform the required action. Hence, the ideal observer does not trust the agent.

The third and final simulation mirrors TACU and its proof more closely. In Simulation 3, if the system cannot prove for any action that a_1 believes a_2 will perform



```
{:name         "Simulation 1"
 :assumptions {;; Agent a2 believes that a1 desires to have block b3 on top of b1
               C1 (Believes! a2 (Desires! a1 (holds (on-top-of b3 b1) t3)))

               C2 (Knows! a2 (Knows! a1
                                      (necessary
                                        (remove b2 b1)
                                        (on-top-of b3 b1))))

               C3 (Controls a1 a2)

               C4 (requests a1 a2 (remove b2 b1) (on-top-of b3 b1))

               NON_AUTONOMOUS_AGENT
                  (forall (?agent1 ?agent2 ?goal ?action ?time)
                          (if (and
                                (Controls ?agent2 ?agent1)
                                (Believes! ?agent1
                                           (Desires! ?agent2 (holds ?goal ?time)))
                                (Believes! ?agent1
                                           (necessary ?action ?goal))
                                (requests ?agent2 ?agent1 ?action ?goal))
                            (perform ?agent1 ?action ?goal)))




               ;; If the system can prove that a2 will perform the action for goal G;
               ;; it can trust the agent.
               TRUST
                  (if
                    (CAN_PROVE! (perform a2 (remove b2 b1) (on-top-of b3 b1)) )
                    (trust a2 (remove b2 b1) (on-top-of b3 b1)))

               }
 :goal    (trust a2 (remove b2 b1) (on-top-of b3 b1))}
```

Fig. 17.6 Simulation 1

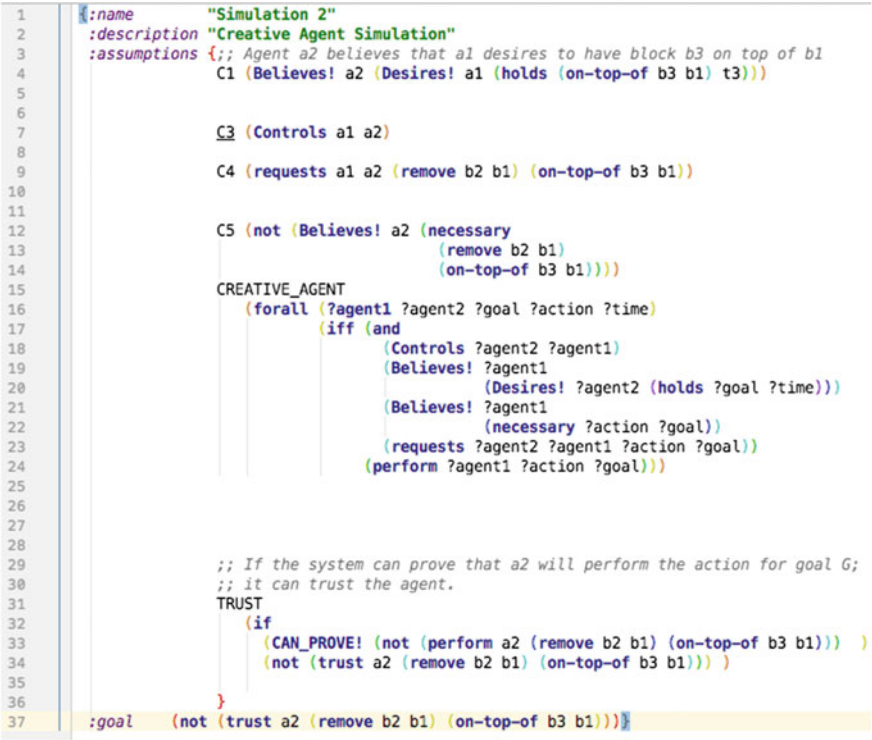
it, and that a_2 will perform that action, then the system cannot trust agent a_2.[16] Next steps along this line, soon to come, include demonstrating these simulations in embodied robots, in real time, with a physicalized Blocks World in our lab.[17]

[16] ShadowProver proves all three problems in around 30 s on a 2011 MacBook Pro with 2.3 GHz Intel Core i7 and 8 GB memory. To obtain a copy of ShadowProver, please contact either of the authors. The simulation input files are available at:

(i) https://gist.github.com/naveensundarg/5b2efebb0aac2f2055fe80012115f195;
(ii) https://gist.github.com/naveensundarg/5f3234f0b93a0a8a34235f5886b225d7; and
(iii) https://gist.github.com/naveensundarg/d061a91f9d966d3cb07c03768b867042

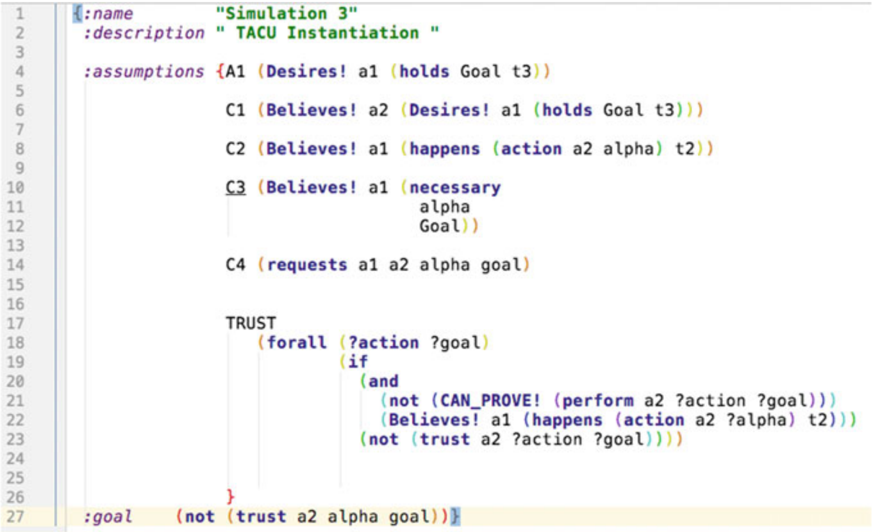
.

[17] We said above that Blocks World is a "time-honored" environment. This is indeed true. In this context, it's important to know that we are only using Blocks World as a convenient venue for making our points clearer and more vivid than they would be if we left things in merely paper-and-pencil form. Hence we are not concerned with displaying raw capability in Blocks World per se. That said, ShadowProver is certainly capable of subsuming landmark achievements in Block's World, such as multi-agent planning [32], and planning via theorem proving [27]. See also Section "Conclusion" for a more recent discussion of social planning in the blocks world.



```
{:name         "Simulation 2"
 :description "Creative Agent Simulation"
 :assumptions {;; Agent a2 believes that a1 desires to have block b3 on top of b1
               C1 (Believes! a2 (Desires! a1 (holds (on-top-of b3 b1) t3)))


               C3 (Controls a1 a2)

               C4 (requests a1 a2 (remove b2 b1) (on-top-of b3 b1))


               C5 (not (Believes! a2 (necessary
                                       (remove b2 b1)
                                       (on-top-of b3 b1))))
               CREATIVE_AGENT
                  (forall (?agent1 ?agent2 ?goal ?action ?time)
                          (iff (and
                                 (Controls ?agent2 ?agent1)
                                 (Believes! ?agent1
                                            (Desires! ?agent2 (holds ?goal ?time)))
                                 (Believes! ?agent1
                                            (necessary ?action ?goal))
                                 (requests ?agent2 ?agent1 ?action ?goal))
                               (perform ?agent1 ?action ?goal)))




               ;; If the system can prove that a2 will perform the action for goal G;
               ;; it can trust the agent.
               TRUST
                  (if
                    (CAN_PROVE! (not (perform a2 (remove b2 b1) (on-top-of b3 b1)))  )
                    (not (trust a2 (remove b2 b1) (on-top-of b3 b1))) )

               }
 :goal    (not (trust a2 (remove b2 b1) (on-top-of b3 b1)))}
```

Fig. 17.7 Simulation 2



```
{:name          "Simulation 3"
 :description " TACU Instantiation "

 :assumptions {A1 (Desires! a1 (holds Goal t3))

               C1 (Believes! a2 (Desires! a1 (holds Goal t3)))

               C2 (Believes! a1 (happens (action a2 alpha) t2))

               C3 (Believes! a1 (necessary
                                  alpha
                                  Goal))

               C4 (requests a1 a2 alpha goal)


               TRUST
                  (forall (?action ?goal)
                          (if
                            (and
                              (not (CAN_PROVE! (perform a2 ?action ?goal)))
                              (Believes! a1 (happens (action a2 ?alpha) t2)))
                            (not (trust a2 ?action ?goal))))


               }
 :goal    (not (trust a2 alpha goal))}
```

Fig. 17.8 Simulation 3

17.5 Toward the Needed Engineering

The chief purpose of the present chapter has been to present the general proposition that supports an affirmative reply to the question that is the chapter's title, and to make a case, albeit a gentle, circumspect one, for its plausibility. We consider this purpose to have been met by way of the foregoing. We end by making two rather obvious points, and reacting to each.

First, TACU is of course enabled by a number of specific assumptions, some of which will be regarded as idiosyncratic by other thinkers; indeed, we anticipate that some readers, outright skeptics, will see both PACU and TACU (and the ingredients used to prove the latter, e.g. TRANS) as flat-out *ad hoc*, despite the fact that both are rooted in the human psyche. For example, there are no doubt some forms of creativity that are radically different than ToM-creativity, and which therefore block the reasoning needed to obtain TACU. (We confess to being unaware of forms of creativity that in no way entail a concept of general "cognitive surprise" on the part of audiences that behold the fruit of such creativity, but at the same time it may well be that we are either under-informed or insufficiently imaginative.) The same can of course be said for our particular regimentation of autonomy. (On the other hand, our oracle-based formalization of autonomy, like ToM-creativity, seems to us to be a pretty decent stand-in for the kernel of *any* fleshed-out form of autonomy.) In reaction, we say that our work can best be viewed as an invitation to others to investigate whether background PACU carries over to alternative formal frameworks.[18] We look forward to attempts on the part of others to either sculpt from the rough-hewn PACU and its empirical support in the human sphere formal propositions that improve upon or perhaps mark outright rejection of elements of TACU, or to go in radically different formal directions than the one we have propaedeutically pursued herein.

The second concluding point is that *if in fact*, as we believe, the background PACU is reflective of a deep, underlying conceptual "flow" from autonomy (our A) and creativity (our C) to untrustworthiness (our U), in which case alternative formal frameworks,[19] once developed, would present counterparts to TACU, then clearly some engineering will be necessary in the future to protect humans from the relevant class of artificial agents: viz. the class of agents that are A and C, and which we wish to enlist in collaborative situations to our benefit.

[18] While we leave discussion of the issue aside as outside our scope herein, we suspect it's worth remembering that some approaches to AI (e.g., ones based exclusively or primarily on such techniques as "deep learning" or reinforcement learning) would presumably by their very nature *guarantee* that the artificial agents yielded by these approaches are U.

[19] While clearly we see dangers in the mixture of autonomy and creativity, which is our focus in the present chapter, if that mixture is expanded to include emotions (to make an "expanded mixture"), the situation is presumably all the *more* worrisome. We leave this remark at the level of the suggestive, but since our cognitive calculi have been used to formalize some of the dominant theories of emotions in cognitive science (including, specifically, the OCC theory itself), it would not be difficult to move from vague worry to more precise treatment of the expanded mixture.

Though the formalism that we have used to state our principle and theorem is explicity logicist, we note that the form of the underlying AI system is not relevant to our theorem. Future explorations of this thread of research can look at more specific AI formalisms such as the AIXI formalism (see Sect. 1.3) and state similar but more specific theorems. For instance, *goal reasoning systems* are systems that can reason over their goals and come up with new goals for a variety of reasons (see Sect. 3.7). Johnson et al. discuss in Sect. 3.7 that trust in such situations must also include trust in the system's ability to reason over goals. We assert that this adds support to our contention that PACU is reflective of a deep, underlying conceptual "flow" from autonomy (our A) and creativity (our C) to untrustworthiness (our U).

If we assume that the future will bring not only artificial agents that are A and C, *but also powerful as well*, the resulting U in these agents is a most unsettling prospect. Our view is that while TACU, as expressed and proved, is by definition idiosyncratic (not everyone in the AI world pursues logicist AI, and not everyone who does uses our cognitive calculi), it is symptomatic of a fundamental vulnerability afflicting the human race as time marches on, and the A and C in AI agents continues to increase in tandem with an increase in the power of these agents.[20]

So what should be done now to ensure that such a prospect is controlled to humanity's benefit? The answer, in a nutshell, is that ethical and legal control must be in force that allows autonomy and creativity in AI systems (since it seems both undeniable and universally agreed that both A and C in intelligent machines has the potential to bring about a lot of good, even in mundane and easy domains like self-driving vehicles) to be developed without endangering humanity.[21] The alert and observant reader will have noticed that $\mathcal{D}^e\mathcal{CEC}$ includes an obligation operator **O** (see again Figs. 17.1 and 17.2), and it would need to be used to express binding principles that say that violating the desires of humans under certain circumstances is strictly forbidden (i.e. it ought/**O** to be that no machine violates the desires of humans in these circumstances). For how to do this (using the very same cognitive calculus, $\mathcal{D}^e\mathcal{CEC}$, used in our three simulations), put in broad strokes, see for instance [6, 29] in our own case,[22] and the work of others who, fearing the sting of future intelligent but immoral machines, also seek answers in computational logic (e.g. [36]).

[20] For a preliminary formalization of the concept of power in an autonomous and creative artificial agent, see [31].

[21] While discussion, even in compressed form, is outside the scope of the present chapter, we would be remiss if we didn't mention that what appears to be needed is engineering that permits creativity in autonomous AI, while at the same time ensuring that this AI technology pursues the goal of sustaining trust in it on the part of humans. Such "trust-aware" machines would have not only ToM-creativity, but, if you will, "ToM prudence."

[22] For more detailed use, and technical presentation, of a cognitive calculus that is only a slightly different dialect than $\mathcal{D}^e\mathcal{CEC}$, see [10]. The results given there are now greatly improved performance-wise by the use of ShadowProver.

References

1. D. Ariely, *The (Honest) Truth About Dishonesty: How We Lie to Everyone — Especially Ourselves* (Harper, New York, NY, 2013). (This is a Kindle ebook)
2. K. Arkoudas, S. Bringsjord, Metareasoning for multi-agent epistemic logics, in *Proceedings of the Fifth International Conference on Computational Logic in Multi-Agent Systems (CLIMA 2004)* (Lisbon, Portugal, September 2004), pp. 50–65
3. K. Arkoudas, S. Bringsjord, Propositional attitudes and causation. Int. J. Softw. Inf. **3**(1), 47–65 (2009)
4. J. Barwise, J. Etchemendy, *Hyperproof* (CSLI, Stanford, CA, 1994)
5. M. Boden, *The Creative Mind: Myths and Mechanisms* (Basic Books, New York, NY, 1991)
6. S. Bringsjord, K. Arkoudas, P. Bello, Toward a general logicist methodology for engineering ethically correct robots. IEEE Intell. Syst. **21**(4), 38–44 (2006)
7. S. Bringsjord, D. Ferrucci, *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine* (Lawrence Erlbaum, Mahwah, NJ, 2000)
8. S. Bringsjord, N.S. Govindarajulu, Given the web, what is intelligence, really? Metaphilosophy **43**(4), 361–532 (2012). (This URL is to a preprint of the paper)
9. S. Bringsjord, N.S. Govindarajulu, S. Ellis, E. McCarty, J. Licato, Nuclear deterrence and the logic of deliberative mindreading. Cogn. Syst. Res. **28**, 20–43 (2014)
10. S. Bringsjord, N.S. Govindarajulu, D. Thero, M. Si, Akratic robots and the computational logic thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)* (Chicago, IL, 2014), pp. 22–29. IEEE Catalog Number: CFP14ETI-POD. Papers from the *Proceedings* can be downloaded from IEEE at http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6883275
11. S. Bringsjord, J. Licato, N. Govindarajulu, R. Ghosh, A. Sen, Real robots that pass tests of self-consciousness, in *Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)* (New York, NY, 2015), pp. 498–504. (IEEE. This URL goes to a preprint of the paper)
12. S. Bringsjord, Declarative/logic-based cognitive modeling, in *The Handbook of Computational Psychology*, ed. by R. Sun (Cambridge University Press, Cambridge, UK, 2008), pp. 127–169
13. S. Bringsjord, Meeting Floridi's challenge to artificial intelligence from the knowledge-game test for self-consciousness. Metaphilosophy **41**(3), 292–312 (2010)
14. S. Bringsjord, N.S. Govindarajulu, Toward a modern geography of minds, machines, and math, in *Philosophy and Theory of Artificial Intelligence, vol. 5, Studies in Applied Philosophy, Epistemology and Rational Ethics*, ed. by V.C. Müller (Springer, New York, NY, 2013), pp. 151–165
15. S. Bringsjord, N.S. Govindarajulu, J. Licato, A. Sen, J. Johnson, A. Bringsjord, J. Taylor, On logicist agent-based economics, in *Proceedings of Artificial Economics 2015 (AE 2015)*, (University of Porto, Porto, Portugal, 2015)
16. S. Bringsjord, A. Sen, On creative self-driving cars: hire the computational logicians, fast. Appl. Artif. Intell. **30**, 758–786 (2016). (The URL here goes only to an uncorrected preprint)
17. S. Bringsjord, J. Taylor, A. Shilliday, M. Clark, K. Arkoudas, Slate: an argument-centered intelligent assistant to human reasoners, in *Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)*, ed. by F. Grasso, N. Green, R. Kibble, C. Reed (University of Patras, Patras, Greece, 21 July 2008), pp. 1–10
18. C. Castelfranchi, R. Falcone, Social trust: a cognitive approach, in *Trust and Deception in Virtual Societies*, ed. by C. Castelfranchi, Y.H. Tan (The Netherlands, Kluwer, Dordrecht, 2001), pp. 55–90
19. R. Chisholm, *Theory of Knowledge* (Prentice-Hall, Englewood Cliffs, NJ, 1966)
20. D. Cope, *Computer Models of Muscial Creativity* (MIT Press, Cambridge, MA, 2005)

21. M. Davis, R. Sigal, E. Weyuker, *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science* (Academic Press, New York, NY, 1994). (This is the second edition, which added Sigal as a co-author)
22. M. Davis, E. Weyuker, *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science* (Academic Press, New York, NY, 1983). (This is the first edition)
23. U. Eco, *The Role of the Reader: Explorations in the Semiotics of Texts* (Indiana University Press, Bloomington, IN, 1979)
24. S. Ellis, A. Haig, N.S. Govindarajulu, S. Bringsjord, J. Valerio, J. Braasch, P. Oliveros, Handle: engineering artificial musical creativity at the 'trickery' level, in *Computational Creativity Research: Towards Creative Machines*, ed. by T. Besold, M. Schorlemmer, A. Smaill (Atlantis/Springer, Paris, France, 2015), pp. 285–308. This is Volume 7 in *Atlantis Thinking Machines*, ed. by Kühnbergwer (Kai-Uwe of the University of Osnabrück, Germany)
25. L. Floridi, Consciousness, agents and the knowledge game. Mind. Mach. **15**(3–4), 415–444 (2005)
26. N. Francez, R. Dyckhoff, Proof-theoretic semantics for a natural language fragment. Linguist. Philos. **33**, 447–477 (2010)
27. M. Genesereth, N. Nilsson, *Logical Foundations of Artificial Intelligence* (Morgan Kaufmann, Los Altos, CA, 1987)
28. G. Gentzen, Investigations into logical deduction, in *The Collected Papers of Gerhard Gentzen*, ed. by M.E. Szabo (North-Holland, Amsterday, The Netherlands, 1935), pp. 68–131. (This is an English version of the well-known 1935 German version)
29. N.S. Govindarajulu, S. Bringsjord, Ethical regulation of robots must be embedded in their operating systems, in *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, ed. by R. Trappl (Switzerland, Springer, Basel, 2015), pp. 85–100
30. J. Johnson, N.S. Govindarajulu, S. Bringsjord, A three-pronged simonesque approach to modeling and simulation in deviant 'bi-pay' auctions, and beyond. Mind Soc. **13**(1), 59–82 (2014)
31. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus, and Giroux, New York, NY, 2013)
32. K. Konolige, N. Nilsson, Multi-agent planning systems, in *Proceedings of Robo-Philosophy 2016*, Proceedings of AAAI–1980 (AAAI, Stanford, CA, 1980), pp. 138–142
33. G. Kreisel, A survey of proof theory II, in *Proceedings of the Second Scandinavian Logic Symposium*, ed. by J.E. Renstad (North-Holland, Amsterdam, The Netherlands, 1971), pp. 109–170
34. J. Licato, N.S. Govindarajulu, S. Bringsjord, M. Pomeranz, L. Gittelson, Analogico-deductive generation of gödel's first incompleteness theorem from the liar paradox, in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI–13)*, ed. by F. Rossi (Morgan Kaufmann, Beijing, China, 2013), pp. 1004–1009. Proceedings are available online at http://ijcai.org/papers13/contents.php. The direct URL provided below is to a preprint. The published version is available at http://ijcai.org/papers13/Papers/IJCAI13-153.pdf
35. N. Marton, J. Licato, S. Bringsjord, Creating and reasoning over scene descriptions in a physically realistic simulation, in *Proceedings of the 2015 Spring Simulation Multi-Conference* (2015)
36. L.M. Pereira, A. Saptawijaya, *Programming Machine Ethics* (Springer, Basel, Switzerland, 2016). (This book is in Springer's SAPERE series, Vol. 26)
37. J. Perry, The problem of the essential indexical. Nous **13**, 3–22 (1979)
38. Dag Prawitz, The philosophical position of proof theory, in *Contemporary Philosophy in Scandinavia*, ed. by R.E. Olson, A.M. Paul (Johns Hopkins Press, Baltimore, MD, 1972), pp. 123–134
39. D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind? Behav. Brain Sci. **4**, 515–526 (1978)
40. A.S. Rao, M.P. Georgeff, Modeling rational agents within a BDI-architecture, in *Proceedings of Knowledge Representation and Reasoning (KR&R-91)*, ed. by R. Fikes, E. Sandewall (Morgan Kaufmann, San Mateo, CA, 1991), pp. 473–484
41. H. Simon, Theories of bounded rationality, in *Decision and Organization*, ed. by C. McGuire, R. Radner (The Netherlands, North-Holland, Amsterdam, 1972), pp. 361–176

42. A. Smith, *Theory of Moral Sentiments* (Oxford University Press, Oxford UK, 1759/1976)
43. R. Soare, *Recursively Enumerable Sets and Degrees* (Springer-Verlag, New York, NY, 1980)
44. A. Turing, Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)