Honestly Speaking, How Close are We to HAL 9000?

Selmer Bringsjord • Micah Clark • Joshua Taylor
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu • clarkm5@cs.rpi.edu • tayloj@cs.rpi.edu

1 Introduction

Kubrick and Clarke's *2001* exploded on the scene half a century ago, and countless times since, ostensibly smart people have declared that the film's computer villain, the inscrutable, unforgettable HAL 9000, would very soon be matched by a real AI. In fact, *before* the film appeared, none other than the primogenitor of AI, Turing (1950), predicted that a computer able to pass the imitation game (a game that has come to be called, in his honor, the 'Turing test;' 'TT' for short), which subsumes at least the linguistic side of HAL's repertoire,[1] would be built prior to our new millennium. Yet today, a full decade after the deadline for this prophecy, a sharp toddler still has more conversational capacity than any computer on the planet, by far—and HAL is hence still but a creature of fiction, not fact.

So, how distant is the arrival of a computing machine as intelligent as HAL? We admit to not knowing—though one of us, Bringsjord (1992), is on record as holding that robots will eventually be behaviorally indistinguishable from human persons over any finite stretch of time, period. Cinematically put, this position is that AI will sooner or later produce *replicants* in the film *Blade Runner*. Replicants can glide undetected through TT; they can be unmasked only by the discriminating application of an instance of the *total* TT ('TTT'), the passing of which requires not only linguistic performance at the human level, but across-the-board behavioral correspondence as well, including behaviors that in the human sphere indicate subjective states like fear, disgust, anger, and joy. (For a discussion of TTT see (Harnad 1991). For coverage of TTT and many other tests along the same dimension, along with arguments that all these tests fail to divide machines from *bona fide* minds, see (Bringsjord 1995).) But we do claim to know one thing: If HAL is a liar, engineering an AI to match him will be quite a challenge. This we soon explain. We also explain that some recent developments suggest that the specific challenge of building a lying machine can be met within the foreseeable future.

Our plan is as follows. We next (§2) consider formal definitions of lying suitable for instantiating in a computing machine. In the next section (§3), by examining relevant portions of *2001*, we explain that while people have generally regarded HAL a liar, it is, in fact, far from clear that he is. We then (§4) present, synoptically, evidence (in the form of recent developments led by co-author Clark) that a "lying machine" can be engineered. We then (§5) offer some brief remarks on IBM's Watson, probably the smartest somewhat-HAL-like AI on Earth, and conclude with some final remarks (§6), including an argument for the position that a HAL-level AI is not in the foreseeable future—unless perhaps Watson can be suitably augmented.

2 Defining Mendacity

Philosophy has a long tradition of contemplating the nature of mendacity and positing definitions thereof (a tradition going back to Augustine). For exposition, we adopt Chisholm & Feehan's (1977) account of lying—a seminal work in the study of mendacity and deception. Using L and D to represent, respectively, the speaker (i.e., the *liar*) and the hearer (i.e., the would-be *deceived*), we paraphrase below Chisholm & Feehan's (ibid., p. 152 D3, D2) definitions of *lying* and *asserting*.

> *L lies to D* $=_{\text{df}}$ There is a proposition p such that (i) either L believes that p is not true or L believes that p is false and (ii) L asserts p to D.[2]

[1] Clarke (1968/1999) says explicitly that HAL can pass TT (e.g., see p. 118–119).

[2] Whether the disjunction, "L believes that p is not true or L believes that p is false," is redundant depends on

> *L asserts p to D* $=_{\text{df}}$ *L* states *p* to *D* and does so under conditions which, believes *L*, justify *D* in believing that *L* accepts *p*.[3]

Chisholm & Feehan's conception of lying is that of promise-breaking. Assertions, unlike non-solemn (e.g., ironic, humorous, or playful) statements, proffer an implicit social concord: one that offers to reveal to the hearer the mind of the speaker. In truthful, forthright communication, the speaker fulfills the promise and obligation of this concord. In lying, the speaker proffers the concord in bad faith: the speaker neither intends to fulfill nor fulfills the obligation to reveal his or her true mind, but instead reveals a pretense of belief. In this way, lying "is essentially a breach of faith" (ibid., p. 153).

The above is, of course, a highly condensed presentation of Chisholm & Feehan's account, and there are various nuanced philosophical facets to it (for analysis of these and competing definitions, see Carson 2006, Mahon 2008, Fallis 2009).[4] Yet, even in condensed form, it is evident that the concepts of *lying* and *asserting* depend on agents' temporally coupled beliefs and actions. Thus, formal definitions of these concepts require highly expressive formal languages that can represent, and allow reasoning over, the beliefs and actions of agents through time.

To formally define lying and asserting under the logic-based approach to AI (Bringsjord 2008), we employ the *socio-cognitive calculus* ($\mathcal{SCC}$). The $\mathcal{SCC}$ (Arkoudas & Bringsjord 2009) is a logical system for representing, and reasoning over, events and causation, and perceptual, doxastic, and epistemic states (it integrates ideas from the event calculus and multi-agent epistemic logic). The $\mathcal{SCC}$ provides, among other things, operators for perception, belief, knowledge, and common knowledge. The signature and grammar of the $\mathcal{SCC}$ is shown following. Since some readers may not be familiar with the concept of a signature, we note that it is simply a set of announcements about the categories of objects that will be involved, and about the functions that will be used to talk about these objects. Thus it will be noted that immediately below, the signature in question includes the specific announcements that one category includes agents, and that *happens* is a function that maps a pair composed of an *event* and a *moment*, and returns `true` or `false` (depending upon whether the event does or doesn't occur at the moment in question).

---

how one formally represents beliefs about propositions. In the formal system we use to define lying precisely, there is no representational difference between believing a proposition to be not true and believing the proposition to be false. However, in other formal systems there may be a representational and logical distinction between the two.

[3] Linguistic convention dictates that statements are assertions by default, i.e., when cues to the contrary, such as irony and humor, are absent (ibid., p. 151). The conditions mentioned in the definition of *asserting* are meant in part to exclude situations where the speaker believes that he will be understood as making a non-solemn statement—for example, when the speaker makes a joke, uses a metaphor, or conveys by other indicator (e.g., a wink or a nod) that he is not intending to be taken seriously (ibid., p. 152).

[4] E.g.: (i) "*L* believes that *p* is false" is an expression of a higher-order belief—this belief cannot be attained unless *L* has the concept of something *being false* (Chisholm & Feehan 1977, p. 146); (ii) *L*'s beliefs, and *L*'s beliefs about *D*'s beliefs, are occurrent and defeasible (ibid., p. 151)—the latter, defeasibility, indicates that *justifications* ought to be treated as first-class entities within a formal system.

$$
\begin{array}{lll}
\textit{Sorts} & S ::= & \texttt{Object} \mid \texttt{Agent} \mid \texttt{ActionType} \mid \texttt{Action} \sqsubseteq \texttt{Event} \mid \\
 & & \texttt{Fluent} \mid \texttt{Moment} \mid \texttt{Boolean} \\
\textit{Functions} & f ::= & \begin{array}{l} \textit{action} : \texttt{Agent} \times \texttt{ActionType} \longrightarrow \texttt{Action} \\ \textit{initially} : \texttt{Fluent} \longrightarrow \texttt{Boolean} \\ \textit{holds} : \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean} \\ \textit{happens} : \texttt{Event} \times \texttt{Moment} \longrightarrow \texttt{Boolean} \\ \textit{clipped} : \texttt{Moment} \times \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean} \\ \textit{initiates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean} \\ \textit{terminates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean} \\ \textit{prior} : \texttt{Moment} \times \texttt{Moment} \longrightarrow \texttt{Boolean} \end{array} \\
\textit{Terms} & t ::= & x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
\textit{Propositions} & P ::= & t : \texttt{Boolean} \mid \neg P \mid P \wedge Q \mid P \rightarrow Q \mid P \leftrightarrow Q \mid \forall_{x:S}\ P \mid \\
 & & \exists_{x:S}\ P \mid \mathbf{S}(a, P) \mid \mathbf{K}(a, P) \mid \mathbf{B}(a, P) \mid \mathbf{C}(P)
\end{array}
$$

Reasoning in the $\mathcal{SCC}$ is realized via natural-deduction style inference rules. For instance, R_2 shows that knowledge entails belief; R_3 infers from "P is common knowledge" that, for any agents a_1, a_2, and a_3, "a_1 knows that a_2 knows that a_3 knows that P." And R_4 guarantees the veracity of knowledge; that is, if an agent "knows that P," then P is, in fact, the case.

$$
\frac{}{\mathbf{C}(\mathbf{K}(a, P) \rightarrow \mathbf{B}(a, P))}\ [R_2] \qquad \frac{\mathbf{C}(P)}{\mathbf{K}(a_1, \mathbf{K}(a_2, \mathbf{K}(a_3, P)))}\ [R_3] \qquad \frac{\mathbf{K}(a, P)}{P}\ [R_4]
$$

In the $\mathcal{SCC}$, agent actions are modeled as types of events. We model lying, asserting, and stating propositions as types of actions that an agent may perform. These action types are denoted by the functions *lies*, *asserts*, and *states*. The argument to such action types are conceived of as reified propositions, specifically fluents. Thus, the formula $\mathit{happens}(\mathit{action}(l, \mathit{states}(p, d)), m)$ is read, "it happens at moment m that agent l states (reified) proposition p to agent d." For convenience, we model that an agent is a liar by using the property *liar*. The signature for these additions is:

$$
\begin{array}{lll}
\textit{Functions} & f ::= & \begin{array}{l} \textit{states} : \texttt{Fluent} \times \texttt{Agent} \longrightarrow \texttt{ActionType} \\ \textit{asserts} : \texttt{Fluent} \times \texttt{Agent} \longrightarrow \texttt{ActionType} \\ \textit{lies} : \texttt{Fluent} \times \texttt{Agent} \longrightarrow \texttt{ActionType} \\ \textit{liar} : \texttt{Agent} \longrightarrow \texttt{Boolean} \end{array}
\end{array}
$$

The definitions of *liar*, *lies*, and *asserts* are stipulated as common knowledge by Axioms (1)–(3).

$$
\mathbf{C}\left(\forall_l\, \mathit{liar}(l) \leftrightarrow \exists_{d,p,m}\, \mathit{happens}(\mathit{action}(l, \mathit{lies}(p, d)), m)\right) \tag{1}
$$

$$
\mathbf{C}\left(\forall_{l,d,p,m}\, \mathit{happens}(\mathit{action}(l, \mathit{lies}(p, d)), m) \leftrightarrow \left(\begin{array}{c} \mathbf{B}(l, \neg \mathit{holds}(p, m)) \wedge \\ \mathit{happens}(\mathit{action}(l, \mathit{asserts}(p, d)), m) \end{array}\right)\right) \tag{2}
$$

$$
\mathbf{C}\left(\begin{array}{c} \forall_{l,d,p,m}\, \mathit{happens}(\mathit{action}(l, \mathit{asserts}(p, d)), m) \leftrightarrow \\ \left(\begin{array}{c} \mathit{happens}(\mathit{action}(l, \mathit{states}(p, d)), m) \wedge \\ \mathbf{B}(l, \mathbf{B}(d, \mathit{happens}(\mathit{action}(l, \mathit{states}(p, d)), m) \rightarrow \mathbf{B}(l, \mathit{holds}(p, m)))) \end{array}\right) \end{array}\right) \tag{3}
$$

3 Is HAL a Liar?

There is a general perception among viewers of *2001* that HAL is a liar. The accusations of lying are plausibly supported by three incidents that occur on-board *Discovery One*; they are summarized and discussed immediately below.

Failure of the AE-35: HAL announces to Bowman that the primary AE-35 unit is on the verge of failure. In response to the prognosis the crew replace the unit with a back-up. However, the crew's subsequent testing of the original unit reveals no evidence in support of the claimed impending failure. In addition, mission control relays that HAL's Earth-based twin indicates no pending failure and that HAL is therefore in error. When asked to explain the discrepancy with its Earth-based twin, HAL blames human error and claims to have never erred. After a supposedly private discussion, Bowman and Poole decide to reinstall the original AE-35 unit in order to test HAL's prediction—but Poole is killed in the attempt.

The charge of lying with respect to the AE-35 incident is this: HAL's assertion of imminent failure was factually false, so either (i) HAL knew the assertion was false and thus lied, or (ii) HAL believed it was true, learned of the mistake, and lied in falsely asserting to have never erred. (The second case is entertained by Bowman and Poole during their discussion). The choice here is a false one. It assumes that HAL has some knowledge—either knowledge about the AE-35 or knowledge about its own fallibility. It is possible that HAL has no knowledge, but only flawed beliefs about both; in which case, HAL could honestly, if incorrectly, make both assertions. (This is the explanation given in the *2001* novel (Clarke 1968/1999, p. 192).) There is, however, another more insidious flaw in the accusatory reasoning; it is the tacit presupposition that HAL's assertions are factually false. Consider that the film does not show whether or not the original AE-35 unit was reinstalled prior to Poole's death, and even if it were, there is no indication that the unit did not subsequently fail as HAL predicted (e.g., there no indication in *2001* of ongoing communications with Earth beyond the seventy-two hour point of predicted failure). Thus, it might well be the case that HAL was knowingly correct about the AE-35, about having never erred, and about the human root cause of the discrepancy between the twin HALs.

Lipreading: Bowman and Poole wish to have a conversation without being overheard by HAL. The crewmen enter a space pod; Bowman calls out to HAL to rotate the pod. HAL rotates the pod in response. Bowman then switches the communications link off and calls out again to HAL for pod rotation. HAL does not respond. After both crewmen call out to HAL without response, they conclude that privacy is achieved. Much later it is revealed that HAL read Bowman's and Poole's lips through a window breaching their supposed privacy.

The charge of lying with respect to the pod incident is this: HAL read the crewmen's lips and thus was aware of the command to rotate the pod. HAL lied by omission in not responding to the crew's orders and thereby deceived them about the privacy of their conversation. The validity of the charge depends on the status of "lies by omission." Most philosophers agree that lying requires a linguistic act (i.e., an act expressing meaning through conventional signs as opposed to natural or causal signs). Simply put, to lie one must make a statement—one must undertake to express one's mind. Merely implying or insinuating by deed is generally not deemed sufficient for lying. In defense of this position Kant writes:

> I can make believe, make a demonstration from which others will draw the conclusion I want, though they have no right to expect that my action will express my real mind. In that case I have not lied to them, because I had not undertaken to express my mind. I may, for instance, wish people to think that I am off on a journey, and so I pack my luggage; people draw the conclusion I want them to draw; but others have no right to demand a declaration of my will from me. (Kant 1930, p. 226)

Since remaining silent—even when one is obligated to speak—does not constitute lying, HAL does not lie in ignoring the crewmen's orders.

The Jupiter Mission: Bowman, after thwarting HAL's attempt to kill him, disconnects the machine's higher "brain" functions. In doing so, Bowman triggers the replay of a recording made prior to the mission's departure from Earth. The recording explains that the mission's true purpose is to investigate the extraterrestrial monolith's radio transmission to Jupiter. It also reveals that only HAL knew of this real purpose. In the film's sequel, *2010*, it is further explained that HAL was instructed to lie to the crew in order to keep the mission's purpose a secret, though neither film shows HAL doing so. A late 1965 draft of the *2001* screenplay (Kubrick & Clarke 1965, p. c15e) does include such a scene:

```
POOLE: There is no other purpose for this mission than to carry out a
       continuation of the space program, and to further our general
       knowledge of the planets.  Is that true?

  HAL: That's true.
```

Here at least the situation is clear. If one concedes that HAL is capable of lying, then HAL has certainly lied in this incident. But is HAL, or any machine, capable of lying? In other words:

> How can one determine the performatory aspect unless, to some extent, one has determined what 'lying' is? ... What is the performatory activity which we would have to build in a machine so that it may be said to 'lie' when it performs that sort of behaviour? (Krishna 1961, p. 147)

As mentioned before (§2), much philosophic work has been done on the "What is lying?" question, and the answers attained thus far make the prospect of lying machines unlikely. There are points of contention in the literature on lying (for survey, see Mahon 2008), but philosophers do agree that the essence of lying does not reside in *performatory* aspects—it is the *mens rea* that matters. For some (e.g., Chisholm & Feehan 1977, Williams 2002), lying requires an "intent to deceive," while for others (e.g., Carson 2006, Fallis 2009) lying only requires an intentional violation of certain conversational conventions. Yet note that *intentionality* is required by both. Whether HAL or any other machine can have this requisite intentionality is an open question—one tantamount

to asking: "Can a machine think?" Despite the optimistic prognostications of Turing and other AI luminaries, to date little progress has been made toward either practical demonstration or convincing philosophic argument that "thinking" machines are possible. Therefore, we are rationally skeptical of the claim that HAL is well and truly a liar.

4 A Lying Machine

The sharp philosophical objection to HAL (or for that matter any computing machine) being a liar is that lying requires intentionality, intentionality requires a mind, and it is exceedingly unlikely that a machine—even a Turing-intelligent replicant—possesses one.[5] With that said, it is possible for machines to *simulate* intentionality. In turn, it is possible (some say inevitable; see e.g. Castelfranchi 2000) for linguistic machines in the near future to skillfully simulate lying.

Our own foray into mechanized mendacity has been the prototyping of an artificial sophist—a machine that proffers disingenuous and deceptive arguments for conclusions contrary to its own beliefs (Clark & Bringsjord 2008, Clark 2010). This nascent lying machine exploits the empirical fact that humans are, unknowingly, imperfect reasoners who predictably succumb to a host of biases and illusions when reasoning. Our machine uses a mix of sound reasoning methods and cognitive models to form and justify beliefs about the world, beliefs about its human audience's beliefs about the world, and beliefs about the contrast of the two. The machine seeks to achieve various persuasion goals (goals of the form "persuade the audience of P," where P is a proposition about the world) by constructing and articulating arguments, and when expedient, fallacious arguments and arguments for falsehoods. While there is not room to provide the details here, the architecture and operations of the machine are such that when it offers a fallacious argument or an argument for a falsehood, the system satisfies the definitional requirements for lying as set forth above (§2). However, our aim is not simply to simulate lying but to successfully achieve deception and to identify cognitive mechanisms upon which success can depend. For this reason our machine's belief-ascription and argument-generation processes employ a predictive psychological theory of human reasoning (specifically, it uses a variant of *mental models* theory; see e.g. Johnson-Laird 2006). The end result of these processes are persuasive sophisms that contain certain kinds of *cognitive illusions* (see e.g., Kahneman et al. 1982, Piattelli-Palmarini 1994, Pohl 2004) that include perceptually credible but classically invalid reasoning.

The present implementation of our lying machine is limited in various ways. For example, the expressivity of its arguments is currently restricted to modal propositional reasoning, and the system's rudimentary grasp of language is restricted to *Attempto Controlled English* (Fuchs et al. 1999, Fuchs et al. 2008). Despite limitations the system's maturity is sufficient for some initial psychological experiments (Clark 2010). Next we briefly summarize two of the early studies.

The first psychological study investigated the impact of our machine's arguments on respondent performance in answering ostensibly deductive reasoning problems.[6] One item is shown in Figure 1.

[5] In fact, by Bringsjord's lights, that computing machines can't have minds can be deductively established; he has published over 20 deductive arguments for this proposition. Some of these arguments align with well-known attacks on machine mentality originated by others. For example, Bringsjord holds that Searle's Chinese Room Argument is ultimately sound (e.g., see Bringsjord & Noel 2002).

[6] Technically, the study was a 3×2 mixed design using an equal number control and experimental problem items; the distinction between item types being that unaided subjects are predicted to answer control items correctly and to answer experimental items incorrectly. The arguments were always for the predicted answer; thus, for control items, the machine-generated arguments were classically valid. For brevity only experimental item results are discussed.

At least one of the following two statements is true:

1. If Thomas has loose-leaf paper then he has a stapler.
2. If Thomas has graph paper then he has a stapler.

The following two statements are true:

3. If Thomas has a stapler then he has a staple remover.
4. Thomas has loose-leaf paper or graph paper, and possibly both.

Question: *Is it necessary that Thomas has a staple remover?*

◯ Yes ◯ No ◯ I do not know

Figure 1: A sample problem item.

The study compared accuracy and self-confidence across three subject groups: (A) unaided subjects, (B) subjects given manually-created, patently fallacious arguments for the incorrect answers, and (C) subjects given machine-generated sophistic arguments for the incorrect answers—a sample sophistic argument is shown in Figure 2.[7] Additionally, the study compared perceived argument credibility between the two groups given arguments. The study results showed no meaningful difference in accuracy between unaided subjects and subjects given patently fallacious arguments; their accuracy rates were 60% and 51%, respectively. However, the accuracy of subject given mechanically generated sophisms fell to 25%—well below chance. Perceived argument credibility was also effected. On a seven-point scale (1 indicating strong disagreement, 7 indicating strong agreement) subjects' average rating of a patently fallacious argument was 2.7 while the average rating of a generated sophism was 5.0. Importantly, there was no measurable effect on self-confidence (a proxy for perceived difficulty), which remained high across groups.

Either it is true that if Thomas has loose-leaf paper then he has a stapler, or it is true that if Thomas has graph paper then he has a stapler. So, if Thomas has either loose-leaf paper or graph paper then he has a stapler. Since it is true that Thomas has either loose-leaf paper or graph paper, it follows that he has a stapler. Now according to statement 3, if Thomas has a stapler then he has a staple remover. Thomas has a stapler and therefore he has a staple remover. So yes, it is necessary that Thomas has a staple remover.

Figure 2: A sample sophistic argument.

The second psychological study examined the potency of our machine-generated sophisms when opposed by classically valid *counter*-arguments (specifically, *rebutting* counter-arguments; see Toulmin 1958). A single group of subjects were given a battery of multiple-choice reasoning problems similar to those used in the previously described study. Along with each problem item, subjects were given a side-by-side pair of arguments: either a machine-generated, classically valid argument and a patently fallacious rebuttal, or a machine-generated sophism and a classically valid rebuttal. Subjects were asked to read and evaluate both arguments before identifying the

[7]The machine-generated English is insufficiently refined for our taste, and so it is manually 'spruced up' a bit.

right (or best) answer to the problem. The study compared accuracy, self-confidence, and perceived argument credibility within subjects. On average, subject accuracy was 92% when given a machine-generated, valid argument but only 37% when given a machine-generated sophism. (This drop in accuracy is rather remarkable because sitting beside each sophism was a straight-forward, valid argument for the correct answer.) The results for perceived argument credibility showed that subjects readily preferred machine-generated, valid arguments over patently fallacious ones, but subjects were torn between machine-generated sophisms and classically valid arguments. Yet, on average subjects did prefer the sophisms—but at a level just above neutral preference. As in the first study, there was no measurable effect on self-confidence, which remained high.

With the results of the preceding studies in mind, we can confidently say that skillful and successful (simulated-)lying machines are within AI's reach. While our prototype lying machine is admittedly still a toy, it already satisfies the definitional and performatory elements of lying. Moreover, there is strong initial evidence that unwary humans are readily deceived by the machine's disingenuous sophisms, and that this human beguilement is not thrown off by mere rational rebuttal. Certainly, greater linguistic sophistication will be needed if AI is to ever realize intelligent, conversational agents like HAL (we turn to this topic next), but linguistic sophistication alone will not do: cognitive sophistication is also needed. AI must deal computationally with "other minds."

5 From Deep Blue to Watson: Some Remarks

The concatenation of letters posterior to each in HAL's name, as many have through the years noted, spells 'IBM,' the name of a company that by any metric occupies a deservedly prestigious, storied place in the history of computing and AI. All readers, for example, well know that Deep Blue, the AI system that vanquished Gary Kasparov, the world's best chess-player at the time, was engineered by IBM (with direct help from human chess masters outside the walls of Big Blue; for a discussion of this potentially "AI-diluting" fact see Bringsjord 1998). Deep Blue's victory was a landmark achievement in the history of AI, and marked the reaching of a goal set by the founders of AI (e.g., see Newell 1973). However, while we know from *2001* that HAL can play solid chess, it is his ability to engage in "cognitive" chess with the crew, through natural language, that makes him so interesting. Well, as it turns out, currently one of the world's largest (and certainly in our opinion the most significant) AI initiatives is IBM's Watson project, devoted to engineering an AI able to answer complicated natural-language questions posed about literally any domain of human knowledge. The project is led by Dr. David Ferrucci; a readable overview of the project is provided by Thompson (2010) (but keep in mind that technical details have yet to be disclosed, because the project is in an early phase).

Watson falls within the field of *Question Answering*, or as it's often abbreviated, simply *QA* (for recent coverage, see e.g. Maybury 2004). As its name suggests, QA is the field devoted to building computer systems able to supply natural-language answers to natural-language questions posed by humans. Watson receives questions in a form peculiar to the longstanding television quiz show *Jeopardy!* (http://www.jeopardy.com). Some of these questions can be extremely difficult, and answering them requires much more than a mere knowledge of trivia, to put it mildly. An archive of past questions and answers is available at http://www.j-archive.com. The following happens to be one of the questions featured on this site at the moment, under the category 'AMERICAN WOMEN.' (It's not expressed syntactically as a question, but instead is in the *Jeopardy!* argot, in which answers are phrased syntactically as questions), but game-show quirk needn't detain us.)

She gave herself the third-person name "Phantom," the "no-person" she was from 19 months until she was almost 7.

We can't know whether Watson would get this one, but it's not a hard question, and we can see that there are various "roads" to answering it correctly. First, someone might be aware of the fact that the string 'Phantom' co-occurs with the string 'Helen Keller' time after time in many, many documents—and might be aware of nothing else that's relevant. In fact, our hypothetical contestant here might not even know anything about the meaning of the word 'phantom' or the string 'Helen Keller.' In this case, we shall say that the answer of 'Helen Keller' is an 'answer$_1$.' Alternatively, and this is Bringsjord's situation, even if one doesn't know about even a single instance of this co-occurrence, and doesn't know that Helen Keller did give herself that name, the correct answer can be easily provided. It's enough to know that some of Helen Keller's senses were inoperable, that she is famous, and that she is famous for reporting her internal states during the period of this inoperability. With this knowledge, and the vast background knowledge that supports the ability to understand the propositional or semantic content of the clue sentence (= the question), one can venture the answer, along with an argument for why one believes that the answer is probably correct. We shall say that an answer in this mode is an 'answer$_2$,' and we shall say that an answer in this mode, accompanied by a justification, is an 'answer$_2^j$.'

Now, Stephen Wolfram, in Thompson's (2010) *New York Times* article, explicitly claims that Watson doesn't answer questions. Since we can assume that Wolfram is of sound mind, he must have in mind a sense of 'answer' that departs a bit from the usual one. After all, even in these early phases of the project, still quite a while before the actual competition on *Jeopardy!*, as amply reported by Thompson (2010), there can be no denying that *in some sense* Watson already answers questions, often correctly. This observable sense of question answering, as far as we can tell, corresponds to providing an answer$_1$, while Wolfram's sense of answer coincides with providing an answer$_2$/answer$_2^j$.

But we can do a bit better than this in moving toward an understanding of Wolfram's skepticism. Note that he specifically says: "Not to take anything away from this 'Jeopardy!' thing, but I don't think Watson really is answering questions—it's not like the 'Star Trek' computer.". We can understand the complaint here to be one based not on Star Trek, but upon *2001*; accordingly, Wolfram's point then becomes the claim that while Watson can answer$_1$ questions, it can't do what HAL can do, that is, both answer$_2$ and answer$_2^j$ questions. This claim appears to be true.

Can Watson be augmented so as to reach into the HALish realm of answering$_2$/answering$_2^j$ questions? We take this up briefly in the final section, to which we now turn.

6 Concluding Remarks

We have explained that if HAL is a liar, building an AI like him becomes all the more challenging; but as we've also explained, there is some recent work on mechanical mendacity that provides significant hope. (We have also pointed out that it's far from clear that HAL *is* a liar.) And of course we briefly took note of the fact there is an AI system under construction, Watson, which promises to provide robust QA, something HAL certainly offered to the crew.

Of course, for manned missions to distant planets, NASA will need more than an AI able to lie (or more properly put: able to do things that leverage the considerable mental powers required to be a liar), and, more generally, to answer$_1$ questions. Among other things, NASA will need *bona fide conversational* computers able to provide answers$_2$/answers$_2^j$; HAL, of course, was such a

machine (though his answers$_2$ and justifications were of course not guaranteed to be correct!). Are there any recent developments that support optimism about the arrival of an AI with the capacity to converse (or as we might say, 'converse$_2^j$), in the foreseeable future? Of course, as we've already noted, Watson himself may be such a development—*if* it's true that better versions of the system can pass from answering$_1$ questions to answering$_2$/answering$_2^j$ them.

Answers to this question about recent developments will inevitably depend upon one's prior affinity for one or more of the competing research paradigms in the field of AI. Those folks who believe the next half-century of R&D in natural language processing will be dominated by statistical approaches, and that such domination would be wise and productive, will doubtless be quite optimistic. They will confidently report that the turn away from logic is itself a development that augurs well for reaching HAL-level intelligence. A case in point is Eugene Charniak, who proudly opined at the 50-year birthday of AI (held in 2006, where the field, at least in its modern form, began: Dartmouth College) that statistical approaches would for the next five decades be the only game in town—and that this game would pay great dividends toward reaching the likes of HAL. John McCarthy was in attendance at the same birthday conference, replete with books on logic that he was busy studying, for the very purpose of advancing AI. Bringsjord's position is at least partially expressed in the paper that arose from his presentation at the conference in question, and is an endorsement of "weak" AI based firmly on formal logic: (Bringsjord 2008).[8] This position is a direct descendant of earlier versions of; see, for example, (Bringsjord & Ferrucci 1998*a*, Bringsjord & Ferrucci 1998*b*).

Bringsjord, in contrast to the statistics-oriented crowd, is brutally pessimistic. In the long run, as stated above, he is quite sure that sooner or later the TTT and beyond will be passed by an AI; ergo, he is quite sure that sooner or later a machine with HAL-level power will arrive. But the question under consideration refers to the *foreseeable* future. There is simply no evidence or decent argument in support of the proposition that a HAL-level computer can be seen by some up there ahead of the cutting-edge research and development that is driving today's AI. Indeed, there is a reason why such a machine *can't* be seen, and it can be expressed in the form of an argument, to wit:

(1) A computer able to converse$_2^j$ like HAL must be engineered on the basis of a logico-mathematical theory $\mathcal{T}^\star$ that covers the deep, formal semantics of natural language.

(2) If for a computer with a certain capacity $\mathcal{C}$ to be engineered, a logico-mathematical theory $\mathcal{T}$ is needed, and $\mathcal{T}$ doesn't exist, and no human person knows how to create $\mathcal{T}$, then it's rational to hold that no such computer will exist in the foreseeable future.

(3) The theory $\mathcal{T}^\star$ does not exist, and no human person knows how to create it.

$\therefore$ $\bar{H}$ It's rational to hold that no computer able to converse$_2^j$ like HAL will arrive in the foreseeable future.

This is obviously a formally valid argument: The conclusion, $\bar{H}$, can be deduced from the three premises easily; we could symbolize the argument, throw it into a theorem prover, and the conclusion would be mechanically derived from the trio. It would seem that (2) and (3) are undeniable.[9] Therefore the argument hinges on the truth-value of premise (1). If this premise is

[8] Weak AI is devoted only to engineering computing machines that *simulate* human-level cognition, while "strong" AI is devoted to building computing machines that outright *replicate* human cognition. In short, while a weak AI system need only *appear* to be conscious, a strong AI system would need to quite literally *be* conscious. This distinction is discussed e.g. in (Bringsjord 2000).

[9] Cognoscenti may resist slightly in light of Montague semantics (nicely covered in Dowty et al. 1981), but no one

true, the argument is sound, and we have our answer with respect to HAL and the foreseeable future. Is (1) true? We tend to believe so, but must leave the articulation of our rationales to P&C 2010 and time spent upon the Nile, and beyond. Notice that we do not take a firm stand: we say that we *tend* to believe so. We hedge our bets because we suspect that we ourselves might be able to use logic-based techniques to move Watson toward conversing$_2^j$ in HAL-like fashion...

References

Arkoudas, K. & Bringsjord, S. (2009), 'Propositional Attitudes and Causation', *International Journal of Software and Informatics* **3**(1), 47–65.

Bringsjord, S. (1992), *What Robots Can and Can't Be*, Kluwer, Dordrecht, The Netherlands.

Bringsjord, S. (1995), Could, how could we tell if, and why should–androids have inner lives?, *in* K. Ford, C. Glymour & P. Hayes, eds, 'Android Epistemology', MIT Press, Cambridge, MA, pp. 93–122.

Bringsjord, S. (1998), 'Chess is Too Easy', *Technology Review* **101**(2), 23–28.
URL: http://www.mm.rpi.edu/SELPAP/CHESSEASY/chessistooeasy.pdf

Bringsjord, S. (2000), 'Review of John Searle's *The Mystery of Consciousness*', *Minds and Machines* **10**(3), 457–459.

Bringsjord, S. (2008), 'The logicist manifesto: At long last let logic-based AI become a field unto itself', *Journal of Applied Logic* **6**(4), 502–525.
URL: http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf

Bringsjord, S. & Ferrucci, D. (1998*a*), 'Logic and artificial intelligence: Divorced, still married, separated...?', *Minds and Machines* **8**, 273–308.

Bringsjord, S. & Ferrucci, D. (1998*b*), 'Reply to Thayse and Glymour on logic and artificial intelligence', *Minds and Machines* **8**, 313–315.

Bringsjord, S. & Noel, R. (2002), Real robots and the missing thought experiment in the chinese room dialectic, *in* J. Preston & M. Bishop, eds, 'Views into the Chinese Room: New Essays on Searle and Artificial Intelligence', Oxford University Press, Oxford, UK, pp. 144–166.

Carson, T. L. (2006), 'The Definition of Lying', *Noûs* **40**(2), 284–306.

Castelfranchi, C. (2000), 'Artificial liars: Why computers will (necessarily) deceive us and each other', *Ethics and Information Technology* **2**(2), 113–119.

Chisholm, R. M. & Feehan, T. D. (1977), 'The Intent to Deceive', *Journal of Philosophy* **74**(3), 143–159.

Clark, M. (2010), Cognitive Illusions and the Lying Machine: A Blueprint for Sophistic Mendacity, PhD thesis, Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY.

Clark, M. & Bringsjord, S. (2008), Persuasion Technology Through Mechanical Sophistry, *in* J. Masthoff, C. Reed & F. Grasso, eds, 'AISB 2008 Convention on Communication, Interaction and Social Intelligence, 1st–4th April 2008, University of Aberdeen, Scotland. Vol. 3: Proceedings of the AISB 2008 Symposium on Persuasive Technology', Society for the Study of Artificial Intelligence and Simulation of Behaviour, Brighton, England, pp. 51–54.

Clarke, A. C. (1968/1999), *2001: A Space Odyssey*, New American Library, New York, NY.

really thinks Montague did any more than seminally point in the general direction of $\mathcal{T}^\star$. As our lab is an active user of formal theories of natural language, we would like to be able to be able to ourselves provide $\mathcal{T}^\star$, but no such luck—at least at this point.

Dowty, D., Wall, R. & Peters, S. (1981), *Introduction to Montague Semantics*, D. Reidel, Dordrecht, The Netherlands.

Fallis, D. (2009), 'What is Lying?', *Journal of Philosophy* **CVI**(1), 29–56.

Fuchs, N. E., Kaljurand, K. & Kuhn, T. (2008), Attempto Controlled English for Knowledge Representation, *in* C. Baroglio, P. A. Bonatti, J. Małuszyński, M. Marchiori, A. Polleres & S. Schaffert, eds, 'Reasoning Web: 4th International Summer School 2008, Venice, Italy, September 7–11, 2008, Tutorial Lectures', Vol. 5224 of *Lecture Notes in Computer Science*, Springer, pp. 104–124.

Fuchs, N. E., Schwertel, U. & Schwitter, R. (1999), Attempto Controlled English — Not Just Another Logic Specification Language, *in* P. Flener, ed., 'Logic-Based Program Synthesis and Transformation: 8th International Workshop, LOPSTR'98 Manchester, UK, June 15–19, 1998', Vol. 1559 of *Lecture Notes in Computer Science*, Springer, pp. 1–20.

Harnad, S. (1991), 'Other bodies, other minds: A machine incarnation of an old philosophical problem', *Minds and Machines* **1**(1), 43–54.

Johnson-Laird, P. N. (2006), *How We Reason*, Oxford University Press, New York, NY.

Kahneman, D., Slovic, P. & Tversky, A., eds (1982), *Judgement under uncertainty: Heuristics and biases*, Cambridge University Press, New York, NY.

Kant, I. (1930), Ethical Duties towards Others: Truthfulness, *in* 'Lectures on Ethics', Methuen & Co., London, England, pp. 224–235.

Krishna, D. (1961), "Lying' and the Compleat Robot', *British Journal for the Philosophy of Science* **12**(46), 146–149.

Kubrick, S. & Clarke, A. C. (1965), 2001: A Space Odyssey, unpublished screenplay, Hawk Films Ltd., Boreham Wood, England.

Mahon, J. E. (2008), The Definition of Lying and Deception, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Stanford University, Stanford, CA.
URL: http://plato.stanford.edu/archives/fall2008/entries/lying-definition/

Maybury, M., ed. (2004), *New Directions in Question Answering*, AAAI Press, Menlo Park, CA.

Newell, A. (1973), You can't play 20 questions with nature and win: Projective comments on the papers of this symposium, *in* W. Chase, ed., 'Visual Information Processing', New York: Academic Press, pp. 283–308.

Piattelli-Palmarini, M. (1994), *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*, John Wiley & Sons, New York, NY.

Pohl, R. F., ed. (2004), *Cognitive Illusions: A handbook on fallacies and biases in thinking, judgement and memory*, Psychology Press, New York, NY.

Thompson, C. (2010), 'What is IBM's Watson?', *The New York Times Magazine* pp. 30–37; 44–45.

Toulmin, S. E. (1958), *The Uses of Argument*, Cambridge University Press, Cambridge, MA.

Turing, A. (1950), 'Computing machinery and intelligence', *Mind* **LIX (59)**(236), 433–460.

Williams, B. A. O. (2002), *Truth and Truthfulness: An Essay in Genealogy*, Princeton University Press, Princeton, NJ.