



# Animals, Zombanimals, and the Total Turing Test

## *The Essence of Artificial Intelligence*

SELMER BRINGSJORD, CLARKE CAPORALE, and RON NOEL

Department of Philosophy, Psychology, and Cognitive Science, Rensselaer Polytechnic Institute,  
Troy, NY 12180-3590, U.S.A.

E-mail: selmer@rpi.edu, caporc@prodigy.net, noelr@rpi.edu

(Received 1 June 1999; in final form 15 April 2000)

**Abstract.** Alan Turing devised his famous test (TT) through a slight modification of the parlor game in which a judge tries to ascertain the gender of two people who are only linguistically accessible. Stevan Harnad has introduced the *Total* TT, in which the judge can *look* at the contestants in an attempt to determine which is a robot and which a person. But what if we confront the judge with an *animal*, and a robot striving to pass for one, and then challenge him to peg which is which? Now we can index TTT to a particular animal and its synthetic correlate. We might therefore have  $TTT_{\text{rat}}$ ,  $TTT_{\text{cat}}$ ,  $TTT_{\text{dog}}$ , and so on. These tests, as we explain herein, are a better barometer of artificial intelligence (AI) than Turing's original TT, because AI seems to have ammunition sufficient only to reach the level of artificial animal, not artificial person.

**Key words:** Artificial intelligence, consciousness, persons, robots, zombies

## 1. Introduction

As is well known, Turing (1964) devised his famous test (TT) through a slight modification of the parlor game in which a judge tries to ascertain the gender of two people who are only linguistically accessible. In TT the judge tries to determine – solely by typed queries and corresponding responses: by email, if you will – which of two contestants is a computer and which is a person. But why not modify the parlor game a bit more violently? First, let us follow Stevan Harnad (1991) out of the pen pal mode: let us allow the judge to not only ask the players by email if they can, say, catch a baseball (and if they say they can, what it *feels* like to catch one); we now permit the judge to *look* at them both, *and* throw each a baseball to catch (or not) before his eyes. Now the test is the so-called *Total* Turing Test (TTT); it challenges the judge to determine which is a robot and which a person. But why stop here? Let us confront the judge with an *animal*, and a robot striving to pass for one, and then challenge him to peg which is which. Now we can index TTT to a particular animal and its synthetic correlate. We might therefore have  $TTT_{\text{rat}}$ ,  $TTT_{\text{cat}}$ ,  $TTT_{\text{dog}}$ , and so on. These tests, as we explain herein, are a better barometer of artificial intelligence (AI) than Turing's original TT, because AI seems to have am-

munition sufficient only to reach the level of artificial animal, not artificial person.\* One of us (Bringsjord) has published many formal, abstract arguments designed to show that no computing machine could ever be a person (e.g., Bringsjord and Zenzen, 1997; Bringsjord, 1992, 1997b). This paper is not centered around another such argument. One of us (Bringsjord) has also argued that TT and TTT (or, more precisely, the claims that such tests determine whether or not something can think, or is conscious), are unacceptable (e.g., see Bringsjord, 1995a). This paper is not centered around these arguments either. Rather, this paper is an attempt to place on the table, and partially defend, the view that AI, as concretely practiced, is chained to tools and techniques sufficient to secure zombanimals, but not persons.

## 2. From Zombies to Zombanimals

The TT encapsulates Turing's empiricist vision for AI. In this vision, obscure concepts like thought, consciousness, free will, and creativity are banished – in favor of sober engineering aimed at producing concrete computational artifacts. The problem is that this engineering can aspire only to build artificial agents at the level of a mere animal, not at the level of a person. In order to demonstrate this, as we said, we will need to appeal to zombanimals.

### 2.1. ZOMBIES AND LOGICAL POSSIBILITY

But what are zombanimals? A zombanimal is at bottom a kind of zombie, except it is a zombie *animal*, or as we say for short, a *zombanimal*. We refer of course not to zombies of cinematic fame (as in, e.g., *The Night of the Living Dead*), but rather to *philosopher's* zombies. These specimens have been introduced via thought-experiments that feature beings displaying our externally observable “input/output” behavior without even a shred of underlying phenomenal consciousness.\*\*

Here is one such gedanken-experiment: You are diagnosed with inoperable brain cancer that will inevitably and quickly metastasize. Desperate, you implore a team of neurosurgeons to replace your brain, piecemeal, with silicon chip work-alikes, until there is only silicon inside your refurbished cranium.‡ The procedure

---

\* A number of Alniks are quite literally trying to build artificial persons. Two such people are the philosophers John Pollock and Daniel Dennett. In his last two books, *How to Build a Person* (Pollock, 1989) and *Cognitive Carpentry: A Blueprint for How to Build a Person* (Pollock, 1995), Pollock argues that in the future his OSCAR system will be a full-fledged person. For Dennett, the person-to-be is the robot COG, or a descendant thereof, a being taking shape with Dennett's help at MIT. (Dennett is helping the well-known roboticist Rodney Brooks.) Dennett shares his vision in Dennett (1994).

\*\* Block (1995), in a recent essay on consciousness in *Behavioral and Brain Sciences*, calls this brand of consciousness **P-consciousness**. We will use a number of words and phrases to refer to this type of consciousness: subjective awareness, inner feelings, having a point of view, that which it is like to be something, etc.

‡ The silicon supplantation is elegantly described in Cole and Foelber (1984).

is initiated . . . and the story then continues in a manner that seems to imply the logical possibility of zombies. In Searle's words:

As the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior . . . [You have become blind, but] you hear your voice saying in a way that is completely out of your control, 'I see a red object in front of me.' . . . We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same (Searle, 1992: 66–67).

Lots of people – Searle, Dennett, Chalmers, Bringsjord, etc.\* – have written about zombies. Most of these thinkers agree that such creatures are logically possible, but, with the exception of Bringsjord (and in a restricted sense Chalmers), they hold that zombies are *physically impossible*. That is, most refuse to accept that such surgery can *in fact* be carried out. Here, in a nutshell, is why Bringsjord believes that the surgery is physically possible.\*\*

## 2.2. ARE ZOMBIES PHYSICALLY POSSIBLE?

Recall the original thought-experiment in which your brain was gradually replaced. This thought-experiment, as we are sure you will agree, is rather short on detail – so short that it provides no reason for anyone to believe that the scenario is physically possible. But why couldn't a neuroscience-schooled Kafka write us a detailed, compelling version of this thought-experiment, replete with wonderfully fine-grained revelations about brain surgery and “neurochips”? This detailed version would include key facts concerning the flow of information in natural neural nets. These facts would be captured in formal architectures; and these architectures would in turn be re-instantiated in human-made computer hardware. The story would be accompanied by a plausible semantic account of physical possibility suitably parasitic on the standard semantic account of logical possibility.‡

Let  $\beta$  denote the state of affairs in which your biological brain is replaced with silicon workalikes, and let  $\diamond$  and  $\diamond_p$  denote logical and physical possibility, respectively. The “physicalized” thought-experiment we have in mind, combined with the account of physical possibility, does not merely establish that  $\diamond\diamond_p\beta$ . Such a proposition says that there is a possible world at which  $\beta$  is physically possible – which is verified by imagining a possible world  $w$  in a cluster of worlds  $w_1, \dots, w_n$  comprising those which preserve the laws of nature in  $w$ , where  $\beta$

---

\* For Searle, see Searle (1992). For Dennett, see Dennett (1991, 1993, 1995). For Chalmers, see Chalmers (1996). Bringsjord's main contribution is in “The Zombie Attack on the Computational Conception of Mind” (Bringsjord, 1999). For more from Bringsjord on zombies, see Bringsjord (1995b).

\*\* For more details, see Bringsjord (1999).

‡ For a number of such accounts, see Earman (1986).

is true not only at  $w$ , but at at least one  $w_i$ . Let  $w_\alpha$  be the actual world; let  $W_\alpha^P$  denote the set of worlds preserving the laws of nature in  $w_\alpha$ . The story we imagine Kafka telling stays scrupulously within  $W_\alpha^P$ . Each and every inch of the thought-experiment is to be devised to preserve consistency with neuroscience and neurosurgery specifically, and biology and physics generally. Our approach here is no different than the approach taken to establish that more mundane states of affairs are physically possible. For example, consider a story designed to establish that brain transplantation is physically possible (and not merely that it is logically possible that it is physically possible). Such a story might fix a protagonist whose spinal cord is deteriorating, and would proceed to include a step-by-step description of the surgery involved, each step described to avoid any inconsistency with neuroscience, neurosurgery, etc. It should be easy enough to convince someone, via such a story, that brain transplantation, at  $w_\alpha$ , is physically possible. (It is of course much easier to convince someone that it is logically possible that it is physically possible that Jones' brain is transplanted: one could start by imagining (say) a world whose physical laws allow for body parts to be removed, isolated, and then made contiguous, whereupon the healing and reconstitution happens automatically, in a matter of minutes.)

We can easily do more than express our confidence in Kafka: we can provide an *argument* for  $\diamond_p \beta$  – if Kafka is suitably armed. There are two main components to this argument. The first is a slight modification of a point made recently by Chalmers (1996), namely, when some state of affairs  $\psi$  seems, by all accounts, to be perfectly coherent (as when it is embedded in narrative, e.g.), the burden of proof is on those who would resist the claim that  $\psi$  is logically possible.\* Specifically, those who would resist need to expose some contradiction or incoherence in  $\psi$ . Most philosophers are inclined to agree with Chalmers here. But then the same principle would presumably hold with respect to *physical* possibility: that is, if by all accounts  $\psi$  seems physically possible, then the burden of proof is on those who would resist affirming  $\diamond_p \psi$  to indicate where physical laws are contravened.

The second component in our argument comes courtesy of the fact that  $\beta$  can be modified to yield  $\beta_{NN}$ , where the subscript “NN” indicates that the new situation appeals specifically to artificial neural networks, which are said to correspond to actual flesh-and-blood brains. So what we have in mind for  $\beta_{NN}$  is this: Kafka now really knows not only about brains and therefore natural neural nets; he also knows about *artificial* ones, and he tells us the sad story about your cancer – but he also tells us how the information flow to, through, and out from your neurons and dendrites is mapped, and how this map allows the gradual replacement of natural brain stuff with artificial correlates in flawless, painstaking fashion, so that information flow in the biological substrate is perfectly preserved in the artificial substrate. In addition, as in the original  $\beta$ , your phenomenal consciousness withers away to zero while your observable behavior runs smoothly on. Now it certainly

---

\* Chalmers gives the case of a mile-high unicycle, which certainly seems logically possible. The burden of proof would surely fall on the person claiming that such a thing is logically impossible.

seems that  $\diamond_p \beta_{NN}$ ; and hence by the principle we isolated above with Chalmers' help, the onus is on those who would resist  $\diamond_p \beta_{NN}$ . This would seem to be a *very* heavy burden. What physical laws are violated in the new, more detailed story? We cannot find any.

### 2.3. ARE ZOMBANIMALS HUMANLY POSSIBLE?

Some things that are physically possible are also *humanly* possible. For example, it is physically possible that the laptop on which Selmer is currently working fall from eight feet directly to the ground. This is also humanly possible: Selmer could stand up, hold the laptop over this head, and let it go. Of course, some physically possible states of affairs are not (at the moment, anyway) humanly possible. For example, it is physically possible that a spaceship travel just short of the speed of light – but NASA can build no such thing.

Now, here is a question we posed to ourselves: the surgery in  $\beta_{NN}$ , if we're right, is physically possible; but is it *humanly* possible? The answer is clearly "No." We are relying on the talents of Kafka, and on the fact that artificial neural nets are very much like real neural nets. But no actual neurosurgeon, at least today, could pull such a thing off. So we posed another question: Could we carry out similar surgery *on animals*? Here our response is different: We believe that zombanimals are humanly (and hence physically) possible. This is our position because one of us (Caporale) has successfully approximated the surgery. Why do we say "approximated"? Well, as in  $\beta_{NN}$ , the idea is to study an animal brain, along with associated sensors and effectors, and to create a precise model of how information flows through the sensors, into the brain, and back out to the triggered effectors. But our surgery is a bit different. We do not seek to *supplant* the animal brain; we seek to instantiate the model in new, silicon hardware; in other words, we seek to *duplicate* the biological creature. Let us explain.

### 3. Simple Zombanimals

To ease exposition, our discussion will now be based upon hypothetical biological creatures whose information-processing architectures are transparent; that is, the flow of information into, through, and out of these creatures has by hypothesis been – to use the term used repeatedly above – *mapped*. The "surgery" (or, as we've explained, duplication) carried out to render these architectures in silicon is real, however. It was carried out in the Minds & Machines Laboratory at Rensselaer.

Let us start with a very simple animal. Not a cat or a rat, something simpler. Imagine a simple multi-cellular organism; let us call it a "bloog." When you shine a penlight on a bloog, it propels itself energetically forward. If you follow the bloog as it moves, keeping the penlight on it, it continues ahead rather quickly. If you shut the penlight off, the bloog still moves ahead, but very, very slowly: the bloog is – we say – listless. If you keep the penlight on, but shine it a few inches away and

in front of the bloog (“front of” being identified with the direction in which it is moving), the bloog gradually accelerates in a straight line in the direction it appears to be facing, but then slows down once it is beyond the light. Given our catalogue of bloog behavior, we can set the  $TTT_{\text{bloog}}$ , and can attempt to build a zombanimal to pass this test.

Caporale has succeeded in precisely this attempt. He has played with a bloog for a while with his penlight, and has witnessed the behavior we have just described; he then set to work. He began by scanning the flow of information in a bloog when one is under a microscope. After a bit, he readied his supply of robotics microhardware, and initiated duplication. Now that he is done, he presents you with . . . a creature he calls “V1.”\* V1 is composed of one tiny sensor and one tiny motor, which are connected, and a structure that supports them. The motor is connected to some device which, when driven by the motor, produces locomotion.\*\*

The behavior of V1 is straightforward: the more of the source detected by its sensor, the faster its motor runs. Were Caporale to give you a demo,‡ you would see that if V1 is bathed in light from the penlight, it moves forward energetically. If it then enters a darker area it becomes listless. If it detects a light ahead, it accelerates toward the light, passes through it, and then decelerates. Obviously, V1 is not subjectively aware: V1 is a zombanimal.

Caporale has also built zombanimal correlates to these two biological creatures: a “sneelock” and a “fleelock.” They are larger than bloogs, a slightly different shade of fleshy brown, and behave differently. A fleelock behaves as follows. If you shine a penlight on the surface on which the fleelock is located, just ahead and exactly in front of the organism, it moves directly toward the light and passes through it, at which point, like bloogs, it becomes listless. However, if you shine the penlight ahead of the fleelock, but to the left or right, it turns to avoid the light, and then moves forward slowly; fleelocks generally dislike light. Sneelocks are similar. They too dislike light, but there is a difference: sneelocks are aggressive. This can be shown by shining a penlight ahead of a sneelock (and, again, to the left or right). When one does this, the sneelock turns with increasing rapidity toward the light, and moves directly at it, eventually moving frontally into the light to apparently assault it.

---

\* Our simple zombanimals are inspired by the vehicular creatures described in Braitenberg’s *Vehicles: Experiments in Synthetic Psychology*. Our first zombanimal is Braitenberg’s Vehicle 1, or just ‘V1’ for short. Note that turning Vn into real robots is not new. Other people have constructed such robots, and you can even buy some of such beasts “off the shelf.” This isn’t a paper on cutting edge robotics; this is a philosophy paper *informed* by real robotics.

\*\* Figures showing the actual zombanimals described in this paper can be found at <http://www.rpi.edu/~faheyj2/SB/SELPAP/ZOMBANIMALS/zombanimals2.pdf>

‡ As he did when we presented inchoate elements of the present paper: “Zombanimals – with Robots from the Minds & Machines Laboratory.” Annual meeting of the Society for Machines and Mentality, at the annual Eastern Division Meeting of the American Philosophical Association, December 1998, Washington, DC.

Caporale was once again allowed to perform “surgery.” After a bit, he cooked up two new zombanimals, V2a and V2b. Courtesy of micro-sensors and motors, V2a behaves just like a fleelock, V2b just like a sneelock. Both V2a and V2b arise from robotic augmentation of V1 that could not suddenly bestow upon them phenomenal consciousness. Hence V2a and V2b are both, like their predecessor, zombanimals. Were you here in our lab, Caporale could show you V1, V2a, and V2b in action.

### 3.1. FROM SIMPLE TO COMPLEX ZOMBANIMALS

You are doubtless thinking that such organisms as bloogs, sneelocks, and fleelocks are excruciatingly simple. Well, you are right. As we have indicated, they are *simple* zombanimals. But Caporale is just warming up.

Consider an animal that can sense and react to not only light, but temperature, oxygen concentration, and amount of organic matter. This biological creature is called a “multi-moog.” Multi-moogs dislike high temperature, turn away from hot places, dislike light with considerable passion (since it turns toward and apparently attempts to destroy it), and prefers a well-oxygenated environment containing many organic molecules. Caporale has “zombified” a multi-moog; the result is V3c. V3c has four pairs of sensors tuned to light, temperature, oxygen concentration, and amount of organic matter. The first pair of sensors is connected to the micro-motors with uncrossed excitatory connections, the second pair with crossed excitatory connections, and the third and fourth pairs with inhibitory connections. It should be obvious that we have no more reason to suppose that V3c is subjectively aware than we have to suppose its predecessors V2a and V2b are: after all, the robotics work that yields V3c from its predecessors consists in a few more wires here and there, and how could such things suddenly bestow upon their bearer phenomenal consciousness?

Before we conclude this section, a reminder regarding our “TTish” theme: multi-moogs can pass versions of the TTT indexed to animals at this level, that is, they can pass TTT<sub>multi-moog</sub>.

## 4. From Zombanimals to Essence of AI

How powerful is Caporale’s toolkit? In the interests of space we will spare you the incremental specification of increasingly complex zombanimals (and we direct you to Braitenberg, 1984). Suffice it to say that Caporale’s arsenal includes primitive logic circuits to full-blown artificial neural networks, loads of sensors and effectors, and some pretty darn good electrical engineering. (The  $V_n$  zombanimals all have the same core brain, a logic board in use in the Minds & Machines Lab. Caporale will soon move to more sophisticated boards.) Caporale is also able to harness evolutionary techniques in order to craft his zombanimals. He will thus have the powers Dennett believes he and others (e.g., Rodney Brooks) have in their

attempt to build the humanoid robot COG (Dennett, 1994).<sup>\*</sup> In sum, it seems to us that Caporale has a toolkit powerful enough to zombify *all* animals. We offer the following inductive argument for this view.

#### 4.1. THE INDUCTIVE ARGUMENT

To begin, we need to idealize things a bit. Accordingly, suppose that biological animals fall across a spectrum isomorphic to some initial sequence of the natural numbers.<sup>\*\*</sup> The first animal,  $b_1$ , would perhaps be a simple single-cell organism, then we would move on to a more sophisticated single-cell organism  $b_2$ , to . . . , to a multi-cellular organism  $b_k$ , to a more sophisticated multi-cellular organism  $b_{k+1}$ , to . . . , then perhaps up to an ant  $b_{k+p}$ , to . . . , to (say) a parrot, to . . . , and eventually up to the animals that would seem to be just “beneath” human persons: apes and chimps. The entire spectrum would then be

$$b_1, b_2, \dots, b_m = \text{ape.}$$

Now, each  $b_i$  has been produced by evolutionary forces, mutations reinforced by natural selection, from  $b_{i-1}$ ; we write  $\mathcal{E}(b_i) = b_{i+1}$  to indicate that evolution caused a step. With this primitive scheme, it is easy to express an inductive argument suggested by the progression seen above in passing from V1 to V2a to V2b to V3c. The basis clause is the proposition that  $b_1$  can be zombified, that is, that a zombanimal, an artificial, non-conscious correlate to a real animal,  $z_1$ , can be engineered which passes  $\text{TTT}_{b_1}$ ; we denote this by  $\mathcal{Z}(b_1)$ . It would seem that such a proposition must be true. It would also seem to be the case that the induction hypothesis is true: if  $b_i$  can be zombified, then through some rather minor engineering modifications so can  $b_{i+1}$ . (Again, how could these tiny engineering tweaks magically generate subjective awareness?) The general principle from which the induction hypothesis follows is this proposition:

$$(*) \forall x \forall y (\mathcal{E}(x) = y \rightarrow (\mathcal{Z}(x) \rightarrow \mathcal{Z}(y)))$$

By the principle of mathematical induction it of course follows that *all* animals can be zombified.

Perhaps some readers will object to the basis clause, as in something like: “Even for TTT competence, it is not so clear to me as it apparently is to you three that a

<sup>\*</sup> For a description of these forces in the context of building zombanimals like the  $V_n$  creatures, see “Vehicle 6: Selection, the Impersonal Engineer,” in Braitenberg (1984).

<sup>\*\*</sup> The “need to idealize things a bit” may strike some readers as quite a jump, but of course we are not supposing that this ordering *literally* exists. After all, that is why we say it is an *idealization*. However, there does seem to be an unshakable intuition that animals go from simple to increasingly intelligent at least roughly along some continuum. Furthermore, it is perhaps not all that implausible to imagine that we could “operationalize” through tests the continuum we invoke. For example, it is well known that rats can be trained to navigate mazes. It is also well known that chimps have been subjected to all sorts of challenges. Perhaps a battery of tests could be devised to solidify the spectrum we imagine.



simple cell is within the competence of AI as you construe it. In particular, the cell has a continuous dynamics, and there are at least some phenomena that can emerge in such systems that cannot be captured with discrete AI systems. This point would also seem to apply to the induction step, for even if  $b_1$  is accepted at the TTT level, it would not necessarily follow that the induction is valid, since the induction is over tinkering with AI architectures, not tinkering in continuous, interactive and metabolic dynamics.”

The problem with this objection is that if a set of processes can be diagrammed and measured in the manner of biological science, this set can be digitized and zombified. One need only pick up a biology textbook to see that creatures at the level of  $b_1$  have been diagrammed, measured, manipulated, altered, and so on. Do we really think that aspects crucial to such creatures have been left out of the picture, *and* that such aspects will *never* be rendered in a form amenable to the kind of engineering at the heart of zombification? Now it is of course true that if a process is genuinely and irreducibly continuous, it may be analog and chaotic, and may thus exceed the reach of computation, and hence may be a process beyond the reach of at least standard AI, whether of the logicist or connectionist sort (see, e.g., Siegelmann, 1995; Siegelmann and Sontag, 1994). But why would anyone think that  $b_1$  harnesses such processes to get by in the world? Surely the burden of proof is on anyone who thinks this.

We imagine that many at this point will also object as follows: “But Caporale is just one lone engineer operating in one isolated laboratory. How do we know that he and his techniques are not idiosyncratic? You yourselves have admitted that he is not working with real animals, but rather with information processing-based idealizations of real animals. Is *that* not idiosyncratic? And yet you seek to generalize wildly from what Caporale has done!”

The rebuttal to this objection is given in the next section, where we explain that the essence of AI is a set of techniques and formalisms for building zombanimals in a manner that coincides remarkably well with Caporale’s engineering.

#### 4.2. AI AS ZOMBANIMAL CONSTRUCTION

On page 7 of their 1985 *Introduction to Artificial Intelligence*, Charniak and McDermott (1985) write: “The ultimate goal of AI (which we are very far from achieving) is to build a person, or, more humbly, an animal.” That the more humble goal is all that AI can reach is of course our main thesis. When you look in detail at Charniak and McDermott’s book, you see there formalisms and techniques sufficient only for creating artificial animals, not persons – that, at least is our position. Unfortunately, even if we are right, the book is over a decade old: it may have captured all of AI in 1985, but may not encompass all of AI *today*.\*

Fortunately, as the century turns, all of AI has been to an astonishing degree unified around a conception that seems to be coextensive with Caporale’s engin-

---

\* As a matter of fact, C&M’s book does not cover sub-symbolic (e.g., neural net-based) AI.

Table I. Lookup table for TABLE-DRIVEN-AGENT.

Percept	Action
001	"Red"
010	"Green"
100	"Blue"
011	"Yellow"
111	"Black"

earing: the conception of an intelligent agent. The unification has in large part come courtesy of a comprehensive textbook intended to cover literally *all* of AI: Russell and Norvig's (1994) *Artificial Intelligence: A Modern Approach (AIMA)*, the cover of which also displays the phrase "The Intelligent Agent Book." The overall, informal architecture for an intelligent agent is presented in the *AIMA* text. According to this architecture, agents take percepts from the environment, process them in some way that prescribes actions, perform these actions, take in new percepts, and continue in the cycle.\*

In *AIMA*, just as is the case in the ladder-like scheme we introduced above for animals and zombanimals, intelligent agents fall on a spectrum from least intelligent to more intelligent to most intelligent. The least intelligent artificial agent is a "TABLE-DRIVEN-AGENT," the program (in pseudo-code) for which is shown in *AIMA*. Suppose that we have a set of actions each one of which is the utterance of a color name ("Green," "Red," etc.); and suppose that percepts are digital expressions of the color of an object taken in by the sensor of a table-driven agent. Then given Table I our simple intelligent agent, running the program in question, will utter (through a voice synthesizer, assume) "Blue" if its sensor detects 100. Of course, this is a stunningly dim agent. What are smarter ones like?

In *AIMA* we reach artificial agents that might strike some as rather smart when we reach the level of a "knowledge-based" agent. The program for such an agent is detailed in *AIMA*. This program presupposes an agent that has a knowledge-base (*KB*) in which what the agent knows is stored in formulae in the propositional calculus, and the functions

- TELL, which injects sentences (representing facts) into *KB*;
- MAKE-PERCEPT-SENTENCE, which generates a propositional calculus sentence from a percept and the time *t* at which it is experienced; and

\* The cycle here is strikingly similar to the overall architecture of cognition described by Pollock (1995).

- MAKE-ACTION-SENTENCE, which generates a declarative fact (in, again, the propositional calculus) expressing that an action has been taken at some time  $t$ .

which give the agent the capacity to manipulate information in accordance with the propositional calculus. (One step up from such an agent would be a knowledge-based agent able to represent and reason over information expressed in full first-order logic.) A colorful example of such an agent is one clever enough to negotiate the so-called “wumpus world,” a so-called “toy world” well-known in AI. The objective of the agent that finds itself in this world is to find the gold and bring it back without getting killed. Pits are always surrounded on three sides by breezes, the wumpus is always surrounded on three sides by a stench, and the gold glitters in the square in which it is positioned. The agent dies if it enters a square with a pit in it (interpreted as falling into a pit) or a wumpus in it (interpreted as succumbing to an attack by the wumpus). The percepts for the agent are given in the form of quadruples. For example,

(Stench,Breeze,Glitter,None)

means that the agent, in the square in which it is located, perceives a stench, a breeze, a glitter, and no scream. A scream occurs when the agent shoots an arrow that kills the wumpus. There are a number of other details involved, but this is enough to demonstrate how command over the propositional calculus can give an agent a level of intelligence that will allow it to succeed in the wumpus world. For the demonstration, let  $S_{i,j}$  represent the fact that there is a stench in column  $i$  row  $j$ , let  $B_{i,j}$  denote that there is a breeze in column  $i$  row  $j$ , and let  $W_{i,j}$  denote that there is a wumpus in column  $i$  row  $j$ . Suppose now that an agent has the following five facts in its  $KB$ .

1.  $\neg S_{1,1} \wedge \neg S_{2,1} \wedge S_{1,2} \wedge \neg B_{1,1} \wedge B_{2,1} \wedge \neg B_{1,2}$
2.  $\neg S_{1,1} \rightarrow (\neg W_{1,1} \wedge \neg W_{1,2} \wedge \neg W_{2,1})$
3.  $\neg S_{2,1} \rightarrow (\neg W_{1,1} \wedge \neg W_{2,1} \wedge \neg W_{2,2} \wedge \neg W_{3,1})$
4.  $\neg S_{1,2} \rightarrow (\neg W_{1,1} \wedge \neg W_{1,2} \wedge \neg W_{2,2} \wedge \neg W_{1,3})$
5.  $S_{1,2} \rightarrow (W_{1,3} \vee (W_{1,2} \wedge W_{2,2} \wedge W_{1,3}))$

Then in light of the fact that

$$\{1, \dots, 5\} \vdash W_{1,3}$$

in the propositional calculus, the agent can come to know (= come to include in its  $KB$ ) that the wumpus is at location column 1 row 3 – and this sort of knowledge should directly contribute to the agent’s success.

Needless to say, a knowledge-based agent, incarnated in robotic hardware, is a zombie: it has no genuine inner life: there is nothing it is like to be such a thing.

Table II. Master table for incremental progression Without consciousness (Key: **An.** for **Animal**, **Zom.** for **Zomanimal**.)

An.	Zom.	Zom. via AIMA	Relevant TTT
$b_1$	$\mathcal{Z}(b_1) = z_1$	$\mathcal{A}(b_1)$	$z_1$ and $\mathcal{A}(b_1)$ pass $\text{TTT}_{b_1}$
$\mathcal{E}(b_1) = b_2$	$\mathcal{Z}(b_2) = z_2$	$\mathcal{A}(b_2)$	$z_2$ and $\mathcal{A}(b_2)$ pass $\text{TTT}_{b_2}$
$\mathcal{E}(b_2) = b_3$	$\mathcal{Z}(b_3) = z_3$	$\mathcal{A}(b_3)$	$z_3$ and $\mathcal{A}(b_3)$ pass $\text{TTT}_{b_3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{E}(b_n) = b_{n+1}$	$\mathcal{Z}(b_{n+1}) = z_{n+1}$	$\mathcal{A}(b_{n+1})$	$z_{n+1}$ & $\mathcal{A}(b_{n+1})$ pass $\text{TTT}_{b_{n+1}}$

Caporale\* has as a matter of fact built an actual wumpus-world-winning robot. This robot, like the  $\forall n$  creatures seen above, is a zomanimal.

Now someone might respond: “Your wumpus-world robot is no doubt a zombie; no problem there. I agree that it has no genuine inner life. But why do you call it a zomanimal? Your zomanimals were non-conscious robotic duplicates of animals. But to what animal does your wumpus-world robot correspond?”

The answer to this question is that zombification could be carried out via the techniques and formalisms that constitute the agent-based approach preached in *AIMA*. Where  $b_i$  is some animal in the continuum invoked earlier, let  $\mathcal{A}(b_i)$  denote the process of zombification – except now the process uses the programs and formalisms in the agent-based approach, along with requisite robotics. Put schematically, the situation so far can be summed up in the progression shown in Table II, where the items in each row are equivalent with respect to both the information processing that drives them, and the phenomenal consciousness that they wholly lack.

## 5. When the Induction Fails: Personhood

We are persons; so are you. On the other hand, we are also biological creatures; in a real and undeniable sense we are animals: *homo sapiens sapiens*. So how is it that Table II, in reaching to animal  $b_{n+1}$ , fails to reach *us*? When we take biological creatures of greater and greater sophistication, and present them to Caporale and an *AIMA*-based engineer for zombification, why will they eventually succeed when we bring them a rat or parrot or a chimp, but fail when we bring them a person? They will fail because they will be impotent in the face of the properties that distinguish persons. What are these properties? Many philosophers have taken up the challenge of answering this question, but for present purposes it suffices to call

\* With help from Tom Poltrino, the Technical Director of the Minds & Machines Laboratory.

upon an account of personhood offered in Bringsjord (1997a); in fact, it suffices to list here only five of the properties offered in that account, viz.,\*

1. ability to communicate in a language,
2. “free will,”
3. creativity,
4. phenomenal consciousness, and
5. robust abstract reasoning (e.g., ability to create conceptual schemes, and to switch from one to another).

For the sake of argument we are prepared to follow Turing and hold that AI will engineer not only the communicative powers of a parrot and a chimp, but also the linguistic powers of a human person. (This concession requires considerable optimism: The current state-of-the art in AI is unable to create a device with the linguistic capacity of a toddler.) However, it is exceedingly hard to see how each of the four remaining properties – a quartet hereafter denoted by “ $\mathcal{L}_4$ ” – can be reduced to circuits and algorithms.  $\forall n$  creatures do not seem to originate anything; they seem to do just what they have been designed to do. And so it is hard to see how they can originate decisions and actions (“free will”) or artifacts (creativity). At least at present, it is hard to see how phenomenal consciousness can be captured in any third-person scheme whatever (and as many readers will recall, a number of philosophers, e.g., Nagel, have argued that such consciousness can never be captured in such a scheme), let alone in something as austere as what AI engineers work with. And those in AI who seek to model abstract reasoning know well that we have only begun to show how sophisticated abstract reasoning can be cast in well-understood computable logics. For all we know at present, it may be that some of this reasoning is beyond the reach of computation. Certainly such reasoning cannot be cashed out in the vocabulary of *AIMA*, which stays firmly within extensional first-order logic.

Please note that we are not claiming here that AI cannot in principle produce a being that has the properties in  $\mathcal{L}_4$ . Bringsjord has made that claim elsewhere, and has provided formal arguments in support of it.\*\* Our claim here is that when one inspects AI, and the tools and techniques that compose it, it certainly *seems* clear that AI only has enough firepower to produce zombanimals up to the level of, say, a chimp. Reaching persons, from the standpoint of what AI has to show and offer, looks to be unattainable.

---

\* The account is streamlined in the interests of space. For example, because people sleep (and because they can be hypnotized, etc.), a person would be a creature with the *capacity* to have properties like those listed here.

\*\* For example, for sustained arguments that literary creativity is beyond the reach of computation, see Bringsjord and Ferruci (2000). For sustained arguments that robust abstract reasoning is beyond computation, see Bringsjord (1997b). For sustained arguments that “free will” is beyond computation, see “Chapter IX: Free Will” in Bringsjord (1992). For sustained arguments that phenomenal consciousness is beyond computation, see Bringsjord and Zenzen (1997), etc.

## 6. Objections

To this point we have heard only six serious objections to our argument. None of them are even close to being victorious.

### 6.1. OBJECTION 1: “YOU’RE BEGGING THE QUESTION!”

Here is the first objection: “You’re just *asserting* that  $\mathcal{L}_4$  can’t be handled by AI. Why should anyone believe this?”

At this point it will perhaps be helpful if we reiterate that our objective in this paper is not to *prove* that AI is in principle unable to build persons and must therefore rest content with building mere animals. Such proofs are provided elsewhere by Bringsjord (again, e.g., Bringsjord, 1992, 1997b; Bringsjord and Ferruci, 2000; Bringsjord and Zenzen, 1997). They start only with the assumption that AI is inseparably wed to the notion that cognition is computation; they do not look, as we have looked and are looking herein, to the nature of AI *qua* real-life field with real-life practitioners. When you look at real-life AI, whether through *AIMA* or any other text or texts or research papers, you come face to face with the brute fact that there is not even a hint of how, e.g., phenomenal consciousness can be bestowed upon an artificial agent or robot. Moreover, you will not even find vague sketches for how P-consciousness will be captured in the future.\*

### 6.2. OBJECTION 2: “YES, BUT *We* EVOLVED!”

The second objection can be expressed like this: “Your argument flies in the face of the fact that *we* evolved; and we are persons. This means that there existed a creature  $c_n$  such that  $\mathcal{E}(c_n)$  was a person, and a creature  $c_{n-1}$  such that  $\mathcal{E}(c_{n-1}) = c_n$ , and a creature  $c_{n-2}$  such that  $\mathcal{E}(c_{n-2}) = c_{n-1}, \dots$ , and so on, back far enough so that we reach the sequence of animals appearing in your Table II. Since zombification (both Caporale’s brand and the *AIMA* variety) is on your own scheme configured to model (or at least include)  $\mathcal{E}$ , it follows that you are wrong that seemingly innocuous incremental changes cannot secure the properties in  $\mathcal{L}_4$  that for you define personhood.”\*\*

This objection is far from decisive, because it presupposes something that is still an open question: viz., that evolution will not need to be rejected or seriously modified in light of “Wallace’s Paradox.” This paradox takes its name from Alfred Wallace, the co-inventor, with Darwin, of the theory of evolution, and arises from the existence of mental powers (e.g., the ability to reason abstractly and rigorously enough to invent the tensor calculus) which seem to have no explanation from the standpoint of evolution. These powers seem to have nothing to do with problems faced by our evolutionary ancestors in their foraging way of life. Or, to put the

\* The same situation holds with respect to the other members of  $\mathcal{L}_4$ . E.g., *AIMA* has nothing on creativity, nothing at all.

\*\* For more on this objection and related issues, see Bringsjord (1999).

point another way, how and why did evolution yield the brain of a modern-day mathematician in order to maximize survival through efficient hunting and gathering of food? Wallace published a paper on primitive people (with whom he lived most of his life) in which he said:

Natural Selection could only have endowed the savage with a brain a little superior to that of an ape, whereas he actually possesses one but a little inferior to that of the average members of our learned societies.

Darwin saw the problem, admitted that it was serious, but declared that some “future psychology” would solve it.

### 6.3. OBJECTION 3: “BUT ANIMALS NEED P-CONSCIOUSNESS FOR X!”

The third objection runs as follows: “Yes, but P-consciousness is required for animals to do X.”\*

Of course, this is a rather anemic objection if *X* is left blank, but fill it in any way you please, and the objection is fundamentally no better.\*\* This point had been eloquently made by Flanagan and Polger (1995). Adapted to the scheme we have erected in this paper, pick any behavior you like for instantiating *X*, and AI can come up with a way of engineering it without anything like P-consciousness entering the picture. Today’s robots, after all, can see and hear and touch things, and reason and act accordingly. How could P-consciousness be required when we have before us, as AI advances, an existence proof that this is wrong? More concretely, and looking a bit down the road, it is hard to see how anyone will be able to rationally maintain that P-consciousness is required for animal behavior when zombanimal correlates for Alex and Lana are with us (see below).

### 6.4. OBJECTION 4: “BUT COGNITIVE SCIENCE IS DIFFERENT!”

The fourth objection reads like this: “Okay, maybe you’re right about AI. But AI, as you yourselves have made clear above, is by definition an *engineering* endeavor; it is not a science. AI is not charged with the task of explaining the nature of personhood in scientific terms. That task falls to the related field of cognitive science, which takes things well beyond mere information processing to theoretical constructs which in turn move beyond zombies.”

This objection is nothing more than wishful thinking; here is why.

---

\* Note that it is not wise to object that *people* need P-consciousness for some *X*. This is because it is undeniable that at least the higher mammals have phenomenal consciousness, which implies that P-consciousness made its advent in the evolutionary story many, many years before *homo sapiens sapiens*. In light of this, the basic idea underlying Objection 3 must apply to animals, not just people.

\*\* In conversation with Bringsjord, Block offered as a candidate for *X* the ability to see. He had in mind, specifically, the ability of a bird to spot its prey from a distance. Block offered this only as a vague possibility.

Alert readers may recall that we made specific reference above to parrots and chimps/apes in the progression of increasingly complex animals. This was tendentious on our parts. We have in mind a part of the cognitive science literature devoted to uncovering the nature of animal cognition. Let us now focus specifically on the remarkable parrot known as Alex, and the wondrous ape known as Lana.

Take Alex first. Recall Table I, and the artificial agent associated with it. As Pepperberg (the scientist who has studied Alex the most) makes clear, a significant part of Alex's intelligence consists in matching the information processing power of such an artificial agent (Pepperberg and Brezinsky, 1991; Pepperberg, 1992). (The color identification problem is specifically one that Alex handles; see Pepperberg and Brezinsky (1991).) Indeed, the "science" in question consists in demonstrating that Alex is capable of cognition that is perfectly modelled by a lookup-table intelligent agent from the *AIMA* paradigm. Alex is capable of other, more impressive feats. For example, he seems to be able to reason in accordance with a proper fragment of the propositional calculus. But here again the science in question simply consists in showing that Alex's cognition includes information processing firmly within the intelligent agent paradigm: Alex is a biological creature who instantiates elements of the primitive knowledge-based agent we presented above.

What about Lana? She is an ape able to instruct her human guardians that she wants a particular food, a particular drink, and so on. (As most readers will doubtless know, there are a number of apes and chimps who through the years have exhibited similar communicative powers. For an excellent overview and discussion of them (including Lana), see Savage-Rumbaugh et al. (1980).) She does this by showing certain icon-imprinted tokens to her guardians. For example, she had icons used to request that she receive milk and that the window be opened. This communicative behavior coincides exactly to the behavior *AIMA* explains how to engineer (in chapters 22 and 23). In fact, in *AIMA* Russell and Norvig design an artificial language for use by an artificial (knowledge-based) agent able to prevail in the wumpus world. So, once again, cognitive science fails to move beyond AI's "zombanimal" paradigm.

Next, the obvious question: What about people? Is it not true that what cognitive science discovers at the level of people moves beyond the models found in AI? No, this is not true at all. Consider the grandest and greatest cognitive architecture for human cognition to be found in cognitive science: ACT-R (Anderson, 1998). ACT-R is intended by Anderson to mark the fulfillment of Newell's dream of "a unified theory" of all human cognition.\* But the amazing thing is that ACT-R is composed of two elementary formalisms and one overarching algorithm, a trio used routinely in AI (and fully covered in *AIMA*). The first formalism is a frame-based representation system, which is merely another way of expressing facts in first-order logic. The second formalism is a production system, which is merely, again, a system that allows for conditional reasoning in first-order logic. The over-

---

\* Newell expressed his dream for a unified (production system-based) theory for all of human cognition in Newell (1973).



arching algorithm is a scheme for planning less sophisticated than those routinely given to today's primitive robots. Where is the science that takes us beyond AI and zombification to something grander? It is nowhere to be found. It is no surprise, then, that the most recent version of ACT-R, version 4.0, is set out in a book that explains, in painstaking detail, how this architecture is instantiated when humans carry out elementary arithmetic (see chapter 9 of Anderson, 1998). But even Alex is nearly capable of such arithmetical cognition.\*

#### 6.5. OBJECTION 5: "YOU SURREPTITIOUSLY RAISED THE BAR FOR PEOPLE!"

The fifth objection is a much more powerful one than its predecessors, and is a bit more involved. It runs as follows.

"When you talk about how your zombanimals capture animal behavior, you are not talking about animals at all. Braitenberg vehicles are very abstract *descriptions* of animal behavior. It is the simplest, bare bones abstraction that we can think of. The same goes for the wumpus world, which is an abstract, formal description of some hypothetical animal behavior. Even the simplest animals are extremely complex organisms capable of doing all sorts of things, and AI has not even come close at this stage to replicating (in the sense of TTT) the behavior of any animal. At best, you three can create machines that pass the TTT for a Braitenberg vehicle, or a TTT for a wumpus world animal, but this is not the same as passing the TTT for a fish or a parrot. For example, you talk of AI's ability to capture the intelligent behavior of Alex, the parrot. You write that 'Alex is a biological creature who instantiates elements of the primitive knowledge-based agent we presented above.' In other words, you insist that contemporary AI can handle Alex-level intelligence, and that the hard problem is human-level intelligence. But, Alex-level intelligence here is the ability to perform at a highly artificial language task that scientists have devised. What is perhaps much more amazing is that Alex's brethren are intelligent enough to survive for many years in an environment teeming with predator and prey, find a mate, successfully raise offspring, and the like. That is what animals do on a day-to-day basis, and your suggestions aside, AI is not close to being able to successfully capture the actual behavior of actual animals. To put it bluntly, the best we can so far do is create pretty good copies of *fake* animals, such as Braitenberg vehicles. Real animals are beyond our grasp at the moment. (Although, there are folks in artificial life and computational neuroethology who are beginning to take the first steps.)

---

\* What we are touching upon here is the "scalability" of ACT-R, and we should mention that chapter 11 in Anderson (1998) is devoted to providing experimental evidence for the view that ACT-R 4.0 can scale up to explain the cognition involved in human scientific discovery. But subjects in the experiment simply hit keys to manipulate variables in a simulated experiment in order to make progress toward ruling in or out a pre-established hypothesis. In short, the subjects do glorified arithmetic, and nothing more, so it is wildly inaccurate to say that they are engaging in scientific discovery.

This is relevant to the your argument, because when you say that AI fails in its goal of replicating human persons, suddenly the bar is not capturing some abstract description of human behavior, but human behavior in all of its richness and glory. I am left wondering why, if the wumpus world is supposed to be taken seriously as a legitimate model of animal behavior, Eliza does not constitute an equally good model of human behavior? It fully captures a rather impoverished abstraction of human cognition in much the same way that Caporale’s robots fully capture a rather impoverished abstraction of animal behavior.”

For at least three reasons, this objection fails.

First, the objection is double-minded regarding our inductive argument. On the one hand, the objection seems to sidestep this argument completely, but on the other it seems to *affirm* it. Notice that our opponent implicitly concedes that we can create a zombanimal able to pass TTT’s beneath a fish or a parrot. And notice that he admits that “folks in artificial life and computational neuroethology” are climbing on up the ordering we invoked in our inductive argument – presumably toward handling a fish or parrot. So which is it? If in fact AI is gradually working up to building zombanimals able to pass  $TTT_{\text{fish}}$  and  $TTT_{\text{parrot}}$ , the objection fails. If our opponent retorts with clarification to the effect that he mispoke, and that he means to stick to his guns that fish and birds are beyond AI’s reach, his objection will amount to nothing, because the plain fact of the matter is that AI *is* managing to slowly but surely capture the real behavior of real animals.

To see the second problem, suppose we take this objection seriously and take pains to consider the “abstract description of human behavior.” Is it really true, as our opponent opines, that such descriptions, like those seen for Braitenberg vehicles, can be captured computationally? No. In fact, one of us (Bringsjord) has provided many examples of such descriptions that are beyond the reach of computation. One such example is found in Bringsjord (1997b); we have enough space here to just quickly describe the basic idea behind the relevant description – but this quick description should demonstrate that the present objection, at best, is inconclusive.

The description concerns infinitary reasoning, and has been given by Bringsjord to explain the behavior of logicians and mathematicians who prove things in and about *infinitary* logics, such as the logic  $\mathcal{L}_{\omega_1\omega}$ , which will be familiar to most readers of this journal.\*

Here is one simple formula in  $\mathcal{L}_{\omega_1\omega}$  which is such that any interpretation that satisfies it is finite:

$$\bigvee_{n < \omega} \exists x_1 \dots \exists x_n \forall y (y = x_1 \vee \dots \vee y = x_n).$$

This formula is an infinite disjunction; each disjunct has a different value for  $n$ . It is well-known that the proposition captured by this formula cannot be expressed in first-order logic, and that the validities of  $\mathcal{L}_{\omega_1\omega}$  cannot be enumerated on a

\* An introductory treatment can be found in Ebbinghaus et al. (1984).

Turing machine. Since the behavior of some logicians and mathematicians centers around infinitary reasoning that can be accurately described only by formalisms that include such formulas (i.e., formalisms like  $\mathcal{L}_{\omega_1\omega}$ ), Objection 5 fails. (Again, for more, see Bringsjord (1997b) and Bringsjord and Zenzen (2001).)

The third problem infecting Objection 5 is that it contains a hidden premise that may or may not be true, viz., that we generally can, at present, give precise abstract descriptions of the entries in  $\mathcal{L}_4$ . Now, scientists can and do give such descriptions for parrots (whether it is their mating behavior or learned counting skills). But what scientist can, today, give a precise abstract description of the cognitive processes that produced such things as *King Lear*?

#### 6.6. OBJECTION 6: “BUT TURING HIMSELF RESPONDED TO THIS!”

The last objection we consider is this one: “Turing (1964) himself, when defending his test, rebutted, to an appreciable degree, the claim that phenomena on your  $\mathcal{L}_4$  cannot be handled by computational techniques. For example, you put creativity on this list. Well, this was basically the objection given by Lady Lovelace in Turing (1964): she claimed that no computer could ever think for itself. But Turing offered a successful reply to this objection.”

While it is true that Turing summarizes an objection from Lady Lovelace that was based, in some sense, on creativity, Turing’s reply is anemic, as we now explain.

Paraphrased, Lovelace’s objection runs like this:

Computers can’t create anything. For creation requires, minimally, *originating* something. But computers originate nothing; they merely do that which we order them, via programs, to do.\*

How does Turing respond? Not well. Lady Lovelace refers here (in her memoirs) to Babbage’s Analytical Engine, which Turing gladly admits did not have the capacity to, as he puts it, “think for itself.” So Turing concedes that insofar as Lovelace’s argument refers to this device, it goes through. But the property of thinking for itself or of originating something is a property Turing assumes to be possessed by *some* discrete state machines, that is, by some computers – ones that arrived *after* Lovelace passed away. Suppose that  $M$  is such a machine. Turing then points out that the Analytical Engine was actually a *universal* digital computer, so if suitably programmed, it could perfectly simulate  $M$ . But such a simulation would bestow upon the Analytical Engine the ability to originate.

Turing’s reasoning here is amazingly bad, for the simple reason that Lovelace would hardly have accepted the assumption that such an  $M$  exists. What machine did Turing have in mind? What machine fits the bill? He does not tell us, but the

\* Scholars take note: We have paraphrased Lady Lovelace in a way that implies her position to be that computers *only* do that which we order them, via programs, to do. See Turing’s footnote 4 in Turing (1964: 29).

fact is that the best he and his contemporaries had to offer were machines whose crowning achievements were merely arithmetical.

Next, Turing inexplicably recasts Lovelace's argument as one for the proposition that computers do not superficially surprise us (Turing, 1964: 21–22) – and he then relates what he takes to be an immediate refutation, viz., “Machines take me by surprise with great frequency.”

This is a terribly weak response. The surprises Turing has in mind arise because because we're not smart and careful enough, or – if we are talking about rare hardware errors – because sometimes microscopic events unfold in unforeseen ways. The unpredictability in question does *not* result from the fact that the computer system has taken it upon itself to *originate* something.

Things are no different if we consider the specific relationship that impresses Turing, namely, the relationship between programmers and their misbehaving programs. Since the three of us regularly program and regularly *teach* programming, we may not be positioned badly to evaluate this relationship.

Suppose that as part of some larger program *P* we seek to write a simple Lisp function to triple a given natural number by producing the following code:

```
(defun triple (n)
  (* m 3))
```

Now suppose that at the Lisp prompt `>` we type `(triple 6)` and get back 75. (Of course, `triple` would in reality be called by another function, but to ease exposition we can assume that we call it directly.) Obviously, *ceteris paribus*, this will surprise us. What is going on? Well, whereas the argument to the function `triple` is said to be `n` in the argument list in the definition of this function, in the body of the function it is `m`, not `n`, that is multiplied by 3. This slight difference, suppose, was the result of a misplaced keystroke. In addition, though we do not remember doing it, for some (smart, let us assume) reason `m` is elsewhere said to be a global variable whose value is 25.\* It seems incredible that a scenario such as this one is one in which the computer or program in question has originated or created something or thought for itself.

We conclude, then, that Turing provides no reason to think we are wrong that AI will falter when attacking  $\mathcal{L}_4$ .\*\*

## 7. Conclusion: Turing's Mistake

A capacity for (flexible and abstract) thought, consciousness, free will, creativity – Turing's 1964 empiricist vision for AI was to banish talk of such obscure concepts in favor of sober engineering aimed at producing computational artifacts able to

\* Of course, no Lisp programmer would use “m” in this way. Some kind of mnemonic would doubtless be employed.

\*\* Turing (1964) does offer another response to Lady Lovelace, namely, one based on the so-called “child machine,” which acquires its intelligence as it “grows up” in the environment. We have insufficient space to explain why this response fails.

pass TT and (by extension) TTT. If what we have said herein is right, this vision has helped mold AI into a field able to produce artificial animals (zombianimals), but a field that will be paralyzed in the attempt to produce persons. If AI manages to produce artifacts that pass TT and TTT (and that thereby *seem* to be persons), this will only be because the field finds a way to trick human judges into believing, on the basis of the right sort of externally observable behavior, that these artifacts are people: trickery because these artifacts will be zombies, *and* because their behavior will be generated by algorithms that fully capture only animal cognition. In TT such artifacts will *say* that they feel such and such, but they will feel no more than the *Vn* zombianimals displayed above feel. Robots are arriving, and they will climb on up the inductive ladder we have described. But to engineer a person? AI has no tools for that, and Turing, as the years tick by, will, we predict, eventually be heralded as he who grasped the computational nature of animals, but expressed mere religious faith that persons were at bottom by nature the same.

### Acknowledgements

For comments on ancestors of this paper, and in some cases for invaluable conversations related to it, we are indebted to Jim Moor, Jim Fetzer, John Searle, John Haugeland, Bill Rapaport, Larry Hauser, Ken Sterling, Eric Steinhart, Michael Zenzen, Tom Poltrino, and Kelsey Rinella. We are also grateful for the penetrating objections of two anonymous referees.

### References

- Anderson, J.R., 1998, *The Atomic Components of Thought*, Mahwah, NJ: Lawrence Erlbaum.
- Block, N., 1995, "On a confusion about a function of consciousness," *Behavioral and Brain Sciences* **18**, 227–247.
- Braitenberg, V., 1984, *Vehicles: Experiments in Synthetic Psychology*, Cambridge, MA: Bradford Books.
- Bringsjord, S., 1992, *What Robots Can and Can't Be*, Dordrecht: Kluwer Academic Publishers.
- Bringsjord, S., 1995a, "Could, how could we tell if, and why should – androids have inner lives?," pp. 93–122 in *Android Epistemology*, K. Ford, C. Glymour, and P. Hayes (eds.), Cambridge, MA: MIT Press.
- Bringsjord, S., 1995b, "In defense of impenetrable zombies," *Journal of Consciousness Studies* **2**, 348–351.
- Bringsjord, S., 1997a, *Abortion: A Dialogue*, Indianapolis, IN: Hackett.
- Bringsjord, S., 1997b, "An argument for the uncomputability of infinitary mathematical expertise," pp. 475–497 in *Expertise in Context*, P. Feltovich, K. Ford, and P. Hayes (eds.), Menlo Park, CA: AAAI Press.
- Bringsjord, S., 1999, "The zombie attack on the computational conception of mind," *Philosophy and Phenomenological Research* **59.1**, 41–69.
- Bringsjord, S. and Ferrucci, D., 2000, *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Mahwah, NJ: Lawrence Erlbaum.
- Bringsford, S. and Zenzen, M., 1997, "Cognition is not computation: The argument from irreversibility," *Synthese* **113**, 285–320.

- Bringsford, S. and Zenzen, M., 2001, *Superminds: A Defense of Uncomputable Cognition*, Dordrecht: Kluwer Academic Publishers, to appear.
- Chalmers, D., 1996, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford, U.K.: Oxford University Press.
- Charniak, E. and McDermott, D., 1985, *Introduction to Artificial Intelligence*, Reading, MA: Addison-Wesley.
- Cole, D. and Foelber, R., 1984, "Contingent Materialism," *Pacific Philosophical Quarterly* **65**, 74–85.
- Dennett, D., 1991, *Consciousness Explained*, Boston, MA: Little, Brown.
- Dennett, D., 1993, "Review of Searle's *The Rediscovery of the Mind*," *Journal of Philosophy* **90**, 193–205.
- Dennett, D., 1994, "The practical requirements for making a conscious robot," *Philosophical Transactions of the Royal Society of London* **349**, 133–146.
- Dennett, D., 1995, "The unimagined preposterousness of zombies," *Journal of Consciousness Studies* **2**, 322–326.
- Earman, J., 1986, *A Primer on Determinism*, Dordrecht: D. Reidel.
- Ebbinghaus, H.D., Flum, J., and Thomas, W., 1984, *Mathematical Logic*, New York: Springer-Verlag.
- Flanagan, O. and Polger, T., 1995, "Zombies and the function of consciousness," *Journal of Consciousness Studies* **2**, 313–321.
- Harnad, S., 1991, "Other bodies, other minds: A machine incarnation of an old philosophical problem," *Minds and Machines* **1.1**, 43–54. This paper is available online at <ftp://cogsci.ecs.soton.ac.uk/pub/harnad/Harnad/harnad91.otherminds>
- Newell, A., 1973, "Production systems: Models of control structures," pp. 463–526 in *Visual Information Processing*, W. Chase (ed.), New York: Academic Press.
- Pepperberg, I., 1992, "Proficient performance of a conjunctive, recursive task by an African gray parrot (*Psittacus erithacus*)," *Journal of Comparative Psychology* **106**, 295–305.
- Pepperberg, I. and Brezinsky, M., 1991, "Acquisition of a relative class concept by an African gray parrot (*Psittacus erithacus*) discrimination based on relative size," *Psychological Monographs* **105**, 286–294.
- Pollock, J., 1989, *How to Build a Person: A Prolegomenon*, Cambridge, MA: MIT Press.
- Pollock, J., 1995, *Cognitive Carpentry: A Blueprint for How to Build a Person*, Cambridge, MA: MIT Press.
- Russell, S. and Norvig, P., 1994, *Artificial Intelligence: A Modern Approach*, Saddle River, NJ: Prentice Hall.
- Savage-Rumbaugh, E.S., Rumbaugh, D.M., and Boysen, S., 1980, "Do apes use language?," *American Scientist* **68**, 49–61.
- Searle, J., 1992, *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.
- Siegelmann, H., 1995, "Computation beyond the Turing limit," *Science* **268**, 545–548.
- Siegelmann, H. and Sontag, E., 1994, "Analog computation via neural nets," *Theoretical Computer Science* **131**, 331–360.
- Turing, A., 1964, "Computing machinery and intelligence," pp. 4–30 in *Minds and Machines*, A.R. Anderson (ed.), Englewood Cliffs, NJ: Prentice-Hall.