Why Did Evolution Engineer Consciousness?*

Selmer Bringsjord
Dept. of Philosophy, Psychology & Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180 USA
selmer@rpi.edu • http://www.rpi.edu/∼brings

Ron Noel
Dept. of Philosophy, Psychology & Cognitive Science
Rensselaer Polytechnic Institute
Troy, NY 12180 USA
noelr@rpi.edu

David Ferrucci
T. J. Watson Research Center
Yorktown Heights, NY 10598 USA
ferrucci@us.ibm.com

April 1, 2000

1 The Question

You, the two of us, the editor of this volume, Plato, Darwin, our neighbors — we have all not only been conscious, but we have also at some point decided to go on living in large part in order to *continue* to be conscious (of that rich chocolate ice cream, a lover's tender touch, the glorious feeling of "Eureka!" when that theorem is finally cracked[1]). For us, consciousness is, to put it barbarically, a big deal. Is it for evolution? Apparently; after all, we evolved. But

Q1 Why did evolution bother to give us consciousness?

*We are indebted to Stevan Harnad for helpful electro-conversations, and Ned Block for *corporeal* conversations, on some of the issues discussed herein.

[1]Things not necessarily to be ranked in the order listed here.

In this paper we refine this question, and then proceed to give what we see as the only sort of satisfactory answer: one which appeals to the intimate connection between consciousness and creativity. Because we both confess to a deep-seated inclination to abide by the dictum "If you can't build it, you don't understand it," we end by relating our answer to Q1 to two radically different attempts at engineering artificial creativity in our laboratory, the second of which — arguably the more promising one — is evolutionary in nature.

2 Taming the Mongrel

Ned Block (1995) has recently pointed out that the concept of consciousness is a "mongrel" one: the term 'consciousness' connotes different things to different people — sometimes *radically* different things. Accordingly, Block distinguishes between

- phenomenal consciousness (**P-consciousness**)
- access consciousness (**A-consciousness**)
- self-consciousness (**S-consciousness**)
- monitoring consciousness (**M-consciousness**)

This isn't the place to carefully disentangle these four breeds. It will suffice for our purposes if we manage to get a rough-and-ready characterization of Block's quartet on the table, with help from Block himself, and some others.

Block describes the first of these phenomena in Nagelian fashion as follows:

> So how should we point to P-consciousness? Well, one way is via rough synonyms. As I said, P-consciousness is experience. P-conscious properties are experiential properties. P-conscious states are experiential states, that is, a state is P-conscious if it has experiential properties. The totality of the experiential properties of a state are "what it is like" to have it. Moving from synonyms to examples, we have P-conscious states when we see, hear, smell, taste and have pains. P-conscious properties include the experiential properties of sensations, feelings and perceptions, but I would also include thoughts, wants and emotions. (Block 1995, p. 230)

According to this explanation, the list with which we began the paper corresponds to a list of P-conscious states, viz.,

- *savoring the taste of rich chocolate*
- *taking pleasure in a lover's caress*
- *experiencing the joy of cracking a proof*

A-consciousness admits of more precise treatment; Block writes:

> A state is access-conscious (A-conscious) if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous, i.e., poised to be used as a premise in reasoning, and (2) poised for [rational] control of action and (3) poised for rational control of speech. (Block 1995, p. 231)

A-consciousness seems to be a property bound up with information-processing. Indeed, as one of us has explained elsewhere (Bringsjord 1997), it's plausible to regard certain extant, mundane computational artifacts to be bearers of A-consciousness. For example, theorem provers with natural language generation capability would seem to qualify with flying colors. In recent conversation, Block has gladly confessed that computational systems, by his lights, are A-conscious.[2]

S-consciousness is said by Block to mean "the possession of the concept of the self and the ability to use this concept in thinking about oneself" (Block 1995, p. 235). There is a famous family of cases (see e.g. Perry 1979) which seem to capture S-consciousness in gem-like fashion: Suppose that you are sitting in the Collar City Diner looking out the window at the passing traffic, when you notice the reflection of a man sitting alone in the diner — a man in a rumpled tweed jacket who is looking out the window with a blank, doleful expression. On the basis of what you see, you affirm, to put it a bit stiffly, this proposition: "The man with the tweed blazer is looking blankly out a window of the Collar City Diner." But suppose you then suddenly realize that the man in question is *you*. At this point you affirm a *different* proposition, viz., "*I* am looking blankly out a window of the Collar City Diner." In this case we say that the indexical is *essential*; and, following Pollock, we say that beliefs that the second sort of proposition hold are *de se* beliefs. We can then say that an agent having *de se* beliefs, as well as the capacity to reason over them (after your epiphany in the diner you may conclude that you need to stop philosophizing and go home and sleep), enjoys S-consciousness.[3]

Block tells us that M-consciousness corresponds to at least three notions in the literature: inner perception, internal scanning, and so-called "higher order" thought. The third of these has been explicated and defended through the years by David Rosenthal (forthcoming, 1986, 1989, 1990*b*, 1990*a*). According to Rosenthal, a state is conscious (in some for-now generic sense of 'conscious') just in case it is the target of a higher-order thought. Courtesy of (Rosenthal forthcoming), the view can be put in declarative form:

Def 1 s is a conscious mental state at time t for agent $a =_{df}$ s is accompanied at t by a higher-order, noninferential, occurrent, assertoric thought s' for a that a is in s, where s' is conscious or unconscious.[4]

[2]The problem is that probably *any* computational artifact will qualify as A-conscious. We think that does considerable violence to our pre-analytic concept of consciousness, mongrel or not. One of us (Bringsjord) has suggested, accordingly, that all talk of A-consciousness be supplanted with suitably configured constituents from Block's definiens. All of these issues — treated in (Bringsjord 1997) — can be left aside without harming the present enquiry.

[3]As Block points out, there are certain behaviors which seem to suggest that chimps enjoy S-consciousness: When colored spots are painted on the foreheads of anesthetized chimps, and the creatures wake and look in a mirror, they try to wipe the spots off (Povinelli 1997). Whether or not the animals really are self-conscious is beside the point, at least for our purposes. But that certain overt behavior is sometimes taken to be indicative of S-consciousness is relevant to what we are about herein (for reasons to be momentarily seen).

[4]Def 1's time index (which ought, by the way, to be a *double* time index — but that's something that needn't detain us here) is necessary; this is so in light of thought-experiments like the following. Suppose (here, as we ask you to suppose again below) that while reading Tolstoy's *Anna Karenina* you experience the state *feeling for Levin's ambivalence toward Kitty*. Denote this state by s^*; and suppose that I have s^* at 3:05 pm sharp; and suppose also that I continue reading without interruption until 3:30 pm, at which time I put down the novel; and assume, further, that from 3:05:01 — the moment at which Levin and Kitty

Def 1, as the **higher-order theory** of consciousness, is often abbreviated as simply 'HOT.' What sorts of examples conform to HOT? Consider the state *wanting to be fed*. On Rosenthal's view, this state *is* a conscious state — and the reason it is is that it's the target of a higher-order thought, viz., the thought that I want to be fed. Rosenthal's Def 1, of course, leaves open the possibility that the higher-order thought can be itself unconscious.

With Block's quartet characterized, it's time to return to Q1, the question with which we began.

3 The Tough Question

Notice first that we now have a specification of Q1 for each member of Block's quartet. For example, we have

$Q1_S$ Why did evolution bother to give us S-consciousness?

This would work similarly for the other breeds of consciousness, giving us $Q1_M$, $Q1_A$, and $Q1_P$. Next, we direct your attention to a variant of Q1, viz.,

Q2 Why might an AI engineer try to give her artifact consciousness?

as well as the corresponding $Q2_X$, with the subscript set to a member of Block's quartet, and 'consciousness' therein changed to 'X-consciousness.' Now, it seems to us that the following principle is true.

P1 If $Q2_X$ is easily answerable, then so is $Q1_X$.

The rationale behind P1 is straightforward: If the AI engineer has a good reason for giving (or seeking to give) her robot consciousness (a reason, we assume, that relates to the practical matter of being a productive robot: engineers who for emotional reasons want to give robots consciousness are of no interest to us[5]), then there is no reason why evolution couldn't have given *us* consciousness for pretty much the same reason.

The interesting thing is that $Q2_S$, $Q2_M$, and $Q2_A$ *do* appear to be easily answerable. Here are encapsulations of the sorts of answers we have in mind.

For $Q2_M$: It should be wholly uncontroversial that robots could be well-served by a capacity to have higher-order thoughts about the thoughts of other robots and humans. For example, a robot working in a factory could exploit beliefs about the beliefs of the humans it's working

temporarily recede from the narrative — to 3:30 I'm completely absorbed in the tragic romance between Anna and Count Vronsky. Now, if I report at 3:30:35 to a friend, as I sigh and think back now for the first time over the literary terrain I have passed, that I feel for Levin, are we to then say that at 3:30:35 s^*, by virtue of this report and the associated higher-order state targeting s^*, becomes a conscious state? If so, then we give me the power to change the past, something I cannot be given.

[5]This is a bit loose; after all, the engineer could want to make a conscious robot specifically for the purposes of studying consciousness. But we could tighten Q2 to something like

Q2′ Why, specifically, might an AI engineer try to give her artifact consciousness in order to make it more productive?

with. Perhaps the robot needs to move its effectors on the basis of what it believes the human believes she sees in front of her at the moment. For that matter, it's not hard to see that it could be advantageous for a robot to have beliefs about what humans believe about what the robot believes — and so on. Furthermore, it's easy to see that robots could benefit from (correct) beliefs about their own inner states. (Pollock (1995) provides a wonderful discussion of the utility of such robotic beliefs.) And for certain applications they could capitalize on a capacity for beliefs about their beliefs about their inner states. (To pick an arbitrary case, on the way to open a comination-locked door a robot may need to believe that it knows that a memory of the combination is stored in memory.[6])

For $Q2_S$: Even a casual look at the sub-field of planning within AI reveals that a sophisticated robot will need to have a concept of itself, and a way of referring to itself. In other words, a clever robot must have the ability to formulate and reason over *de se* beliefs. In fact, it's easy enough to adapt our diner case from above to make the present point: Suppose that a robot is charged with the task of making sure that patrons leave the Collar City Diner in time for the establishment to close down properly for the night. If, when scanning the diner with its cameras, the robot spots a robot that appears to be simply loitering near closing time (in a future in which the presence of robots is commonplace), it will certainly be helpful if the employed robot can come to realize that the loitering robot is just itself seen in a mirror.

For $Q2_A$: This question can be answered effortlessly. After all, A-consciousness can pretty much be *identified* with information processing. Any reason a roboticist might have for building a robot capable of reasoning and communicating will be a reason for building a robot with A-consciousness. For this reason, page after page of standard textbooks contain answers to $Q2_A$ (see e.g. Russell & Norvig 1994).

Question $Q2_P$, however, is another story. There are at least two ways to see that P-consciousness is quite another animal. The first is to evaluate attempts to reduce P-consciousness to one or more of the other three breeds. One such attempt is Rosenthal's HOT. This theory, as incarnated above in Def 1, didn't take a stand on what breed of consciousness is referred to in the definiendum. Rosenthal, when asked about this, bravely modifies Def 1 to yield this definition:

Def 1_P s is a P-conscious mental state at time t for agent $a =_{df}$ s is accompanied at t by a higher-order, noninferential, occurrent, assertoric thought s' for a that a is in s, where s' may or may not be P-conscious.

Unfortunately, Def 1_P is very implausible. In order to begin to see this, let s'' be one of our paradigmatic P-conscious states from above, say *savoring a spoonful of deep, rich chocolate ice cream.* Since s'' is a P-conscious state, "there is something it's like" to be in it. As Rosenthal admits about states like this one:

> When [such a state as s''] is conscious, there is something it's like for us to be in that state. When it's not conscious, we do not consciously experience any of its qualitative

[6] You may be thinking: Why should a robot need to believe such a thing? Why not just be able to *do* it? After all, a simple photoactic robot need not believe that it knows where the light is. Well, actually, the case we mention here is a classic one in AI. The trick is that unless the robot believes it has the combination to the lock in memory, it is irrational to for it to fire off an elaborate plan to get to the locked door. If the robot doesn't know the combination, getting to the door will have been a waste of time.

> properties; so then there is nothing it's like for us to be in that state. How can we explain this difference? ... How can being in an intentional state, of whatever sort, result in there being something it's like for one to be in a conscious sensory state? (Rosenthal forthcoming, pp. 24–25)

Our question exactly. And Rosenthal's answer? He tells us that there are "factors that help establish the correlation between having HOTs and there being something it's like for one to be in conscious sensory states" (Rosenthal forthcoming, p. 26). These factors, Rosenthal tells us, can be seen in the case of wine tasting:

> Learning new concepts for our experiences of the gustatory and olfactory properties of wines typically leads to our being conscious of more fine-grained differences among the qualities of our sensory states ... Somehow, the new concepts appear to generate new conscious sensory qualities. (Rosenthal forthcoming, p. 27)

But Rosenthal's choice of wine tasting tendentious. In wine tasting there is indeed a connection between HOTs and P-conscious states (the nature of which we don't pretend to grasp). But wine-tasting, as a source of P-consciousness, is unusually "intellectual," and Def 1_{P} must cover all cases — including ones based on less cerebral activities. For example, consider fast downhill skiing. Someone who makes a rapid, "on-the-edge" run from peak to base will have enjoyed an explosion of P-consciousness; such an explosion, after all, will probably be the main reason such an athlete buys expensive equipment and expensive tickets, and braves the cold. But expert downhill skiers, while hurtling down the mountain, surely don't analyze the ins and outs of pole plants on hardpack versus packed powder surfaces, and the fine distinctions between carving a turn at 20 mph versus 27 mph. Fast skiers ski; they plunge down, turn, jump, soar, all at incredible speeds. Now is it really the case, as Def 1_{P} implies, that the myriad P-conscious states $s_1, \ldots, s_n$ generated in a screaming top-to-bottom run are the result of higher-level, noninferential, assertoric, *occurrent* beliefs on the part of a skier k that k is in s_1, that k is in s_2, k is in s_4, ..., k is in s_n? Wine tasters do indeed sit around and say such things as that, "Hmm, I believe this Chardonnay has a bit of a grassy taste, no?" But what racer, streaking over near-ice at 50 mph, ponders thus: "Hmm, with these new parabolic skis, 3 millimeters thinner at the waist, the sensation of this turn is like turning a corner in a fine vintage Porsche". And who would claim that such thinking *results in* that which it's like to plummet downhill?

C	F	P	Y
J	M	B	X
S	G	R	L

Figure 1: Sample 3 × 4 Array for Backward Masking

Def 1_{P} is threatened by phenomena generated not only at ski areas, but in the laboratory as well. We have in mind an argument arising from the phenomenon known as **backward masking**. Using a tachistoscope, psychologists are able to present subjects with a visual stimulus for periods of time on the order of milliseconds (one millisecond is 1/1000th of a second). If a subject is shown a 3 × 4 array of random letters (see Figure 1) for, say, 50 milliseconds (msecs), and is then asked to report the letters seen, accuracy of about 37%

is the norm. In a set of very famous experiments conducted by Sperling (1960), it was discovered that recall could be dramatically increased if a tone sounded after the visual stimulus. Subjects were told that a high tone indicated they should report the top row, a middle tone the middle row, and a low tone the bottom row. After the table above was shown for 50 msec, to be followed by the high tone, recall was 76% for the top row; the same result was obtained for the other two rows. It follows that a remarkable full 76% of the array is available to subjects after it appears. However, if the original visual stimulus is followed immediately thereafter by another *visual* stimulus in the same location (e.g., circles where the letters in the array appeared; see Figure 2), recall is abysmal; the second visual stimulus is said to backward mask the first (the seminal study is provided in Averbach & Coriell 1961). Suppose, then, that a subject is flashed a series of visual patterns p_i, each of which appears for only 5 msec. In such a case, while there is something it is like for the subject to see p_i, it is very doubtful that this is because the subject thinks that she is in p_i. In fact, most models of human cognition on the table today hold that information about p_i never travels "far enough" to become even a potential object of any assertoric thought (Ashcraft 1994).

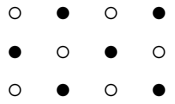

Figure 2: Sample Visual 3 × 4 Array Used as Stimulus in Backward Masking Experiments

So, for these reasons, Def 1_{P} looks to us to be massively implausible. More generally, the point is that it's extraordinarily difficult to reduce P-consciousness to other forms of consciousness.

The *second* way to see that P-consciousness is much more recalcitrant than the other three breeds we have singled out is to slip again into the shoes of the AI engineer. Why would a roboticist strive to give her creation the capacity to experience that which it's like to, say, eat an ice cream cone? It would seem that any reason the robot might have for consuming chocolate fudge swirl in a waffle cone could be a reason devoid of any appeal to P-consciousness. (Perhaps the robot needs cocoa for fuel (other types of energy sources turned out to be a good deal more expensive, assume); but if so, it can be built to seek cocoa out when it observes that its power supply is low — end of story, and no need to appeal to anything as mysterious as subjective awareness.) Evolution *qua* engineer should similarly find P-consciousness to be entirely superfluous.[7] Which is to say that we have moved from Q1 to what we call the "tough" question:

$Q1_{\text{P}}$ Why did evolution bother to give us P-consciousness?

This question can in turn be further refined through "zombification."

[7]John Pollock, whose engineering efforts, he avows, are dedicated to the attempt to literally build an artificial person, holds that emotions are at bottom just timesavers, and that with enough raw computing power, the advantages they confer for us can be given to an "emotionless" AI — as long as the right algorithms are in place. See his discussion of emotions and what he calls **Q&I modules** in (Pollock 1995).

4 Zombifying the Question

In order to zombify the tough question we need to restructure it so that it makes reference to zombies. The zombies we have in mind are *philosophers*' zombies, not those creatures who shuffle about half-dead in the movies.[8] Philosophers' zombies, to use Stevan Harnad's (1995) felicitous phrase, are bodies with "nobody home" inside. Such zombies are characters in a variation arising from a gedanken-experiment lifted directly out of the toolbox most philosophers of mind, today, carry with them on the job: Your brain starts to deteriorate and the doctors replace it, piecemeal, with silicon chip workalikes which flawlessly preserve the "information flow" within it, until there is only silicon inside your refurbished cranium.[9] John Searle (1992) claims that at least three distinct variations arise from this thought-experiment:

V1 The Smooth-as-Silk Variation: The complete silicon replacement of your flesh-and-blood brain works like a charm: same mental life, same sensorimotor capacities, etc.

V2 The Zombie Variation: "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, pp. 66–7).

V3 The Curare Variation: Your body becomes paralyzed and the doctors, to your horror, give you up for dead.[10]

Scenario V2 seems to us to be clearly **logically possible** (a proposition written, using the possibility operator from modal logic, as $\Diamond$V2); that is, V2 seems to us to be a scenario free from contradiction, perfectly coherent and conceivable. After all, Searle could, at the drop of a hat, provide a luxurious novel-length account of the scenario in question (or he could hire someone with the talents of a Kafka to do the job for him).[11]

[8] The zombies of cinematic fame apparently do have real-life correlates created with a mixture of drugs and pre-death burial: see (Davis 1985, Davis 1988).

[9] For example, the toolbox is opened and the silicon supplantation elegantly pulled out in (Cole & Foelber 1984).

[10] This scenario would seem to resemble a real-life phenomenon: the so-called "Locked-In" Syndrome. See (Plum & Posner 1972) (esp. the fascinating description on pages 24-5) for the medical details.

[11] Despite having no such talents, one of us (Bringsjord) usually spends twenty minutes or so telling a relevant short story to students when he presents zombies via V2. In this story, the doomed patient in V2 — Robert — first experiences an unintended movement of his hand, which is interpreted by an onlooker as perfectly natural. After more bodily movements of this sort, an unwanted sentence, to Robert's mounting horror, comes involuntarily up from his voicebox — and is interpreted by an interlocutor as communication from Robert. The story describes how this weird phenomenon intensifies ... and finally approaches Searle's "late stage" description in V2 above. Now someone might say: "Now wait a minute. Internal amazement at dwindling consciousness requires differing *cognition*, a requirement which is altogether incompatible with the preservation (*ex hypothesi*) of identical "information flow". That is, in the absence of an argument to the effect that ordinary cognition (never mind consciousness) fails to supervene on "information flow", V2 is incoherent". The first problem with this objection is that it ignores the ordering of events in the story.

Not everyone sees things the way we do. Daniel Dennett has registered perhaps the loudest and most articulate dissent. In fact, Dennett has produced an argument (based, by the way, on the Rosenthalian definition of M-consciousness discussed above) for $\neg\Diamond$V2 in his recent *Consciousness Explained* (Dennett 1991, pp. 304–313).[12] One of us (Bringsjord) has formalized Dennett's argument (Bringsjord 1999), and found it wanting, but there isn't space here to recapitulate the argument. We don't ask that you regard this attempted refutation to be sound, sight unseen. We *do* ask that, for the sake of argument, you join the many prominent thinkers who affirm the likes of $\Diamond$V2 (e.g., Dretske 1996, Block 1995, Chalmers 1996, Flanagan & Polger 1995, Harnad 1995). Moreover, we ask that you grant that V2 is *physically* possible, that is, that V2, though no doubt monstrously improbable, could come to pass without violating any laws of nature in our world. This seems to us to be a reasonable request to make of you. After all, why couldn't a neuroscience-schooled Kafka write us a detailed, compelling account of V2, replete with wonderfully fine-grained revelations about brain surgery and "neurochips"? Then we have only to change the modal operator to its physics correlate — $\Diamond$ to $\Diamond_p$.[13] Each and every inch of the thought-experiment in question 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's logically possible that it's 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 is physically possible.[14]

This last assertion will no doubt be challenged; we hear some readers saying: "Surely the two of you must be joking. To concede that such neural implantation is physically possible is (debatable) one thing, but to concede that *and* that the result would be a V2-style zombie

Robert, *earlier*, has had his brain supplanted with silicon workalikes — in such a way that all the same algorithms and neural nets are in place, but they are just instantiated in different physical stuff. Then, a bit later, *while these algorithms and nets stay firmly and smoothly in place*, Robert fades away. The second problem with the present objection is that it's a clear *petitio*, for the objection is that absent an argument that consciousness is conceptually distinct from information flow, the thought-experiment fails (is incoherent). But the thought-experiment is designed for the specific purpose of showing that information flow is conceptually distinct from consciousness! If X maintains that, necessarily, if p then q, and Y, in attempt to overthrow X's modal conditional, describes a scenario in which, evidently, p but $\neg q$, it does no good for X to say: "Yeah, but you need to show that p can be present without q". In general, X's only chance is to grapple with the issue in earnest: to show that the thought-experiment is somehow defective, despite appearances to the contrary.

[12]This is an argument on which Dennett has recently placed his chips: In his recent "The Unimagined Preposterousness of Zombies" (1995) Dennett says that the argument in question shows that zombies are not really conceivable.

[13]For cognoscenti: we could then invoke some very plausible semantic account of this formalism suitably parasitic on the standard semantic account of $\Diamond$. For a number of such accounts, see (Earman 1986).

[14]It is of course much easier to convince someone that it's logically possible that it's 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.

is absurd. In any case, if it is 'perfectly reasonable' to allow V2 as a physical possibility, then anything extra about logical possibility is superfluous, since the former entails the latter."

The part of this objection which consists in observing that

$$\Diamond_p \phi \rightarrow \Diamond \phi$$

is certainly correct; this conditional is obvious and well-known. But why is it *absurd* that $\Diamond_p$V2? Isn't the progression to $\Diamond_p$V2 quite sensible? We start with the story (from, e.g., Searle) designed to establish $\Diamond$V2; and then when we look at this story we ask the question: What laws of nature are broken in it? Again, why can't Kafka give us a novel *showing* $\Diamond_p$V2?

Let us make it clear that we can easily do more than express our confidence in Kafka: We can provide an *argument* for $\Diamond \mathrm{V2}_1^{\mathrm{TM}}$ given that 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 ψ seems, by all accounts, to be perfectly coherent, the burden of proof is on those who would resist the claim that ψ is logically possible.[15] Specifically, those who would resist need to expose some contradiction or incoherence in ψ. We think 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 ψ 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 V2 can be modfied to yield $\mathrm{V2}^{\mathrm{NN}}$, where the superscript 'NN' indicates that the new situation appeals to artificial **n**eural **n**etworks, which are said to correspond to actual flesh-and-blood brains.[16] So what

[15] 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. This may be the place to note that Chalmers considers it *obvious* that zombies are both logically and physically possible — though he doesn't think zombies are *naturally* possible. Though we disagree with this position, it would take us too far afield to consider our objections. By the way, Chalmers (1996, pp. 193–200) refutes the only serious argument for the logical impossibility of zombies not mentioned in this paper, one due to Shoemaker (1975).

[16] A quick encapsulation: Artificial neural nets (or as they are often simply called, 'neural nets') are composed of **units** or **nodes** designed to represent neurons, which are connected by **links** designed to represent dendrites, each of which has a numeric **weight**. It is usually assumed that some of the units work in symbiosis with the external environment; these units form the sets of **input** and **output** units. Each unit has a current **activation level**, which is its output, and can compute, based on its inputs and weights on those inputs, its activation level at the next moment in time. This computation is entirely local: a unit takes account of but its neighbors in the net. This local computation is calculated in two stages. First, the **input function**, in_i, gives the weighted sum of the unit's input values, that is, the sum of the input activations multiplied by their weights:

$$in_i = \sum_j W_{ji} a_j.$$

In the second stage, the **activation function**, g, takes the input from the first stage as argument and generates the output, or activation level, a_i:

$$a_i = g(in_i) = g\left(\sum_j W_{ji} aj\right).$$

One common (and confessedly elementary) choice for the activation function (which ususally governs all units in a given net) is the step function, which usually has a threshold t that sees to it that a 1 is output

we have in mind for $V2^{NN}$ is this: Kafka really knows his stuff: he knows not only about natural neural nets, but also about artificial ones, and he tells us the sad story of Smith — who has his neurons and dendrites gradually replaced with artificial correlates in flawless, painstaking fashion, so that information flow in the biological substrate is perfectly preserved in the artificial substrate ... and yet, as in V2, Smith's P-consciousness withers away to zero while observable behavior runs smoothly on. Now it certainly seems that $\Diamond_p V2^{NN}$; and hence by the principle we isolated above with Chalmers' help, the onus is on those who would resist $\Diamond_p V2^{NN}$. This would seem to be a *very* heavy burden. What physical laws are violated in the new story of Smith?

We are now in position to "zombify" $Q1_P$:

$Q1_P^Z$ Why did evolution bother to fashion us, bearers of P-consciousness, rather than zombies, creatures — courtesy of the right sort of information processing working in unison with sensors and effectors — having our behavioral repertoire, but lacking our inner lives?

5 Creativity as an Answer

There are at least three general ways to answer $Q1_P^Z$:

A1 "Look, evolution does allow for outright accidents, so maybe P-consciousness is just an adventitious 'add on' having no survival value 'payoff' — in which case there would be no particular reason why evolution fashioned us rather than zombies."[17]

A2 "The answer is that P-consciousness has some definite function, and though this function can be carried out by mere information processing (suitably symbiotic with the outside environment on the strength of sensors and effectors), evolution took the more interesting route."

A3 "P-consciousness has a definite function, yes, but one arguably not replicable in any system based on the standard information processing built into the gedanken-experiments taken to substantiate $\Diamond$V2."

A1 is really not an answer to $Q1_P^Z$; and as such it's profoundly unsatisfying (if the informal poll we've taken is any indication). It even seems downright bizarre to hold that the phenomenon that makes life worth living (Wouldn't you be depressed upon hearing that starting five minutes from now you would have the inner life of a slab of granite?) is a fluke. The second answer, A2, is given by a respondent who hasn't grasped the problem: After all, if the function of P-consciousness can be carried out by computation, then why didn't

when the input is greater than t, and that 0 is output otherwise. This is supposed to be "brain-like" to some degree, given that 1 represents the firing of a pulse from a neuron through an axon, and 0 represents no firing. As you might imagine, there are many different kinds of neural nets. The main distinction is between **feed-forward** and **recurrent** nets. In feed-forward nets, as their name suggests, links move information in one direction, and there are no cycles; recurrent nets allow for cycling back, and can become rather complicated. Recurrent nets underlie the MONA-LISA system we describe below.

[17] As Ned Block has recently pointed out to one of us (Bringsjord), since at least all mammals are probably P-conscious, the accident would had to have happened quite a while ago.

evolution take the programming route? This question is just $Q1_P^Z$ all over again, so A2 gets us nowhere.[18]

A3 is the answer we favor. This means we have to be prepared to step up to the challenge and show that certain behaviors do correspond to P-consciousness, and that obtaining these behaviors from ordinary computation isn't possible. What behaviors might qualify? In a word: creativity. We conclude this section by providing reason to believe that P-consciousness' role in us is to enable creativity. In the following section, when we discuss our engineering work, we return to the view that creativity requires more than standard information processing.

One of us (Bringsjord: (Bringsjord 1997)) has recently tried to reconstruct a Searlean argument for the view that a — perhaps *the* — function of P-consciousness is to enable creativity. Searle's argument is enthymematic; its key hidden premise is a principle which unpacks the common-sense idea that if the advent of a psychological deficiency coincides with a noteworthy diminution of a person's faculty, then it's a good bet that the diminution is causally linked with the deficiency. With (a slightly more sophisticated version of) this principle (P2), we can produce a Searlean argument that is formally valid in first-order logic. The argument runs as follows.[19]

$\mathbf{A}_1$

P2	If S loses x over an interval of time during which S loses the ability to ϕ, and there are substantive reasons to think x is centrally employed when people ϕ (in part because (i) attempts to replicate ϕ-ing in systems lacking x have failed, and show no appreciable promise of succeeding in the future; and (ii) subjects report that they need x in order to ϕ), then a function of x is to at least facilitate ϕ-ing.	
(1)	S loses x over an interval of time during which S loses the ability to ϕ, ... that they need x in order to ϕ).	
(2)	There is at least a *prima facie* reason to think x is centrally employed when people ϕ (in part because attempts to replicate ϕ-ing in systems lacking x have failed, and show no appreciable promise of succeeding in the future).	
$\therefore$ (3)	A function of x is to facilitate ϕ-ing.	P2, 1, 2

[18]This is eloquently explained by Flanagan & Polger (1995), who explain the some of the functions attributed to P-consciousness can be rendered in information-processing terms.

[19]Some may object to P2 in this way: "*Prima facie*, this is dreadfully implausible, since each (x and ϕ) may be an effect of a cause prior to both. This has a form very similar to: provided there is constant conjunction between x and ϕ, and someone somewhere thinks x is centrally imployed in ϕ-ing, x actually does facilitate ϕ-ing."

Unfortunately, this counter-argument is very weak. The objection is an argument from analogy — one that supposedly holds between defective inferences to causation from mere constant conjunction to the inference P2 promotes. The problem is that the analogy breaks down: In the case of P2, there is more, *much* more, than constant conjunction (or its analogue) to recommend the inference — as is explicitly reflected in P2's antecedent: it makes reference to evidence from reports and from the failure of certain engineering attempts. (Some of the relevant reports are seen in the case of Ibsen. One such report is presented below.)

Victorious instantiations of this schema seem to us to be at hand. (If $x =$ 'P-consciousness,' and $\phi =$ 'write belletristic fiction,' then it turns out that one of us has elsewhere (Bringsjord & Ferrucci 2000) explicitly defended the relevant instantiation. The basic idea underlying the instantiation is that creativity, for example the creativity shown by a great dramatist, requires P-consciousness.[20] The defense capitalizes on P2's parenthetical by including an observation that AI has so far failed to produce creative computer systems.

You may ask, "Yes, but what evidence have you for P2?" We haven't space to include here all of the evidence for this principle. Some of it is empirical (e.g., (Cooper & Shepard 1973)); some of it is "commonsensical." Evidence of the latter sort is obtained by remembering that all of us have experienced unplanned intervals of "automatism." To repeat the familiar example, you're driving late at night on the interstate; you're 27 miles from your exit ... and the next thing you know, after reverie about a research problem snaps to an end, you are but *seventeen* miles from your turnoff. Now, was there anything it was like to drive those ten mysterious miles? If you're like us, the answer is a rather firm "No" (and we daresay the real-life cases are myriad, and not always automotive). Now, why is it that such episodes invariably happen when the ongoing overt behavior is highly routinized? Have you ever had such an episode while your overt behavior involved, say, the writing of sentences for a short story, or the writing of inferences toward the proving of a theorem? These are rhetorical questions only, of course. But surely it's safe to say that P2 is no pushover, and that A_1 constitutes a respectable case for the view that a function of P-consciousness is to enable creative cognition.

6 Engineering Creativity

The foregoing discussion, quite theoretical in nature, is related to two attempts on our part to "engineer creativity." The first attempt is, as we say, a **parameterized** one based in formal logic: a system designed to generate interesting short-short[21] stories with help from pre-set formalizations designed to capture the *essence* of such stories. As we explain, this

[20]Henrik Ibsen wrote:

> I have to have the character in mind through and through, I must penetrate into the last wrinkle of his soul. I always proceed from the individual; the stage setting, the dramatic ensemble, all that comes naturally and does not cause me any worry, as soon as I am certain of the individual in every aspect of his humanity. (reported in (Fjelde 1965), p. xiv)

Ibsen's *modus operandi* is impossible for an agent incapable of P-consciousness. And without something like this *modus operandi* how is one to produce creative literature?

At this point we imagine someone objecting as follows. "The position expressed so far in this paper is at odds with the implied answer to the rhetorical question, Why can't impenetrable zombies write creative literature? Why can't an impenetrable zombie *report* about his *modus operandi* exactly as Ibsen did, and then proceed to write some really great stories? If a V2 zombie is not only logically, but even physically possible, then it is physically possible that Ibsen actually had the neural implant procedure performed on him as a teenager, and no one ever noticed (and, of course no one *could* notice)."

The reply to this objection is simple: Sure, there is a physically possible world w in which Ibsen's output is there but P-consciousness isn't. But the claim we're making, and the one we need, is that internal behavior of the sort Ibsen *actually* engaged in ("looking out through the eyes of his characters") requires P-consciousness.

[21]Our term for stories about the length of Betrayal.1. Stories of this type are discussed in (Bringsjord & Ferrucci 2000).

attempt (at least as it currently stands) seems to be threatened by the apparent fact that the space of all interesting short-short stories cannot be captured in some pre-set formalism — the space may be, in the technical sense of the term, **productive**. (For now — we furnish the technical account below—, understand a productive set to be one whose membership conditions can't be formalized in computational terms.) Our second engineering attempt is evolutionary in nature, and steers clear of any notion that creativity consists in "filling in" the parameters in some pre-defined account of the desired output.

6.1 BRUTUS: A Parameterized Logicist Approach

We are members of the Creative Agents project at Rensselaer,[22] which extends and enhances the Autopoeisis Project. Autopoeisis, launched in 1991 with grants from the Luce Foundation, with subsequent support from Apple Computer and IBM, is devoted to building an artificial storyteller capable of generating "sophisticated fiction." (A snapshot of the project's first stage was provided in (Bringsjord 1992).) Though we confess that no such AI is currently on the horizon (anywhere), the BRUTUS system suggests that the dream driving Autopoeisis may one day arrive. (BRUTUS is the overall system architecture. The first incarnation of that architecture is BRUTUS_1. Details may be found in (Bringsjord & Ferrucci 2000).) Though they aren't exactly Shakespearean, BRUTUS is able to produce stories such as the following one.

BETRAYAL.1

> Dave Atwood loved the university. He loved its ivy-covered clocktowers, its ancient and sturdy brick, and its sun-splashed verdant greens and eager youth. He also loved the fact that the university is free of the stark unforgiving trials of the business world — only this *isn't* a fact: academia has its own tests, and some are as merciless as any in the marketplace. A prime example is the dissertation defense: to earn the PhD, to become a doctor, one must pass an oral examination on one's dissertation. This was a test Professor Edward Hart enjoyed giving.
>
> Dave wanted desperately to be a doctor. But he needed the signatures of three people on the first page of his dissertation, the priceless inscriptions which, together, would certify that he had passed his defense. One of the signatures had to come from Professor Hart, and Hart had often said — to others and to himself — that he was honored to help Dave secure his well-earned dream.
>
> Well before the defense, Dave gave Hart a penultimate copy of his thesis. The professor read it and told Dave that it was absolutely first-rate, and that he would gladly sign it at the defense. They even shook hands in Iron's book-lined office. Dave noticed that Ed's eyes were bright and trustful, and his bearing paternal.
>
> At the defense, Dave thought that he eloquently summarized Chapter 3 of his dissertation. There were two questions, one from Professor Rogers and one from Dr. Teer; Dave answered both, apparently to everyone's satisfaction. There were no further objections.
>
> Professor Rogers signed. He slid the tome to Teer; she too signed, and then slid it in front of Hart. Hart didn't move.

[22] Information can be found at `http://www.rpi.edu/dept/ppcs/MM/c-agents.html`.

"Ed?" Rogers said.

Hart still sat motionless. Dave felt slightly dizzy.

"Ed, are you going to sign?"

Later, Hart sat alone in his office, in his big leather chair, saddened by Dave's failure. He tried to think of ways he could help Dave achieve his dream.

BRUTUS can generate stories like this one because, among other reasons, it "understands" the literary concepts of self-deception and betrayal via formal definitions of these concepts. (The definitions, in their full formal glory, and the rest of BRUTUS' anatomy, are described in (Bringsjord & Ferrucci 2000).) To assimilate these definitions, note that betrayal is at bottom a relation holding between a "betraye**r**" (s_r in the definition) and a "betraye**d** (s_d in the definition)." Then here is a (defective) definition that gives a sense of BRUTUS's "knowledge:"

Agent s_r betrays agent s_d iff there exists some state of affairs p such that

1 s_d wants p to occur;

2 s_r believes that s_d wants p to occur;

3 s_r agrees with s_d that p ought to occur;

4′ there is some action a which s_r performs in the belief that thereby p will *not* occur;

5′ s_r believes that s_d believes that there is some action a which s_r performs in the belief that thereby p *will* occur;

6′ s_d wants that there is some action a which s_r performs in the belief that thereby p *will* occur.

BRUTUS also has knowledge of story structures in the form of story grammars. For example, BETRAYAL.1 conforms to the following story grammar, taken from Thorndyke (Thorndyke 1977). That which flanks '+' comes sequentially; the asterisk indicates indefinite repetition; parentheses enclose that which is optional; brackets attach to mutually exclusive elements.

Rule No.	Rule
(1)	Story → Setting + Theme + Plot + Resolution
(2)	Setting → Characters + Location + Time
(3)	Theme → (Event)* + Goal
(4)	Plot → Episode*
(5)	Episode → Subgoal + Attempt* + Outcome
(6)	Attempt → { Event* ; Episode }
(7)	Outcome → { Event* ; State }
(8)	Resolution → { Event ; State }
(9)	{ Subgoal ; Goal } → Desired State
(10)	{ Characters ; Location ; Time } → State

Is BRUTUS creative? Perhaps not.[23] After all, BRUTUS is capable of generating only a small portion of the space $\mathcal{I}$ of all interesting short-short stories: the formalisms that make up BRUTUS's soul seem to leave out much of this space.[24] Even a future incarnation of BRUTUS, BRUTUS$_n$, that includes knowledge of *all* presently deployed literary concepts (unrequited love, friendship, revenge, etc.), all known story grammars, and so on — even such a system, we suspect, would inevitably fail to capture the essence of $\mathcal{I}$. Our suspicion is based on the intuition that $\mathcal{I}$ is productive, in the technical sense: A set Φ is productive if and only if (i) Φ is classically undecidable (= no program can decide Φ), and (ii) there is a computable

[23]For reasons explained in (Bringsjord & Ferrucci 2000), BRUTUS *does* seem to satisfy the most sophisticated definition of creativity in the literature, one given by (Boden 1995).

[24]It may be thought that brute force can obviously enumerate a superset of $\mathcal{I}$, on the strength of reasoning like this:

> Stories are just strings over some finite alphabet. Given the stories put on display on behalf of BRUTUS.1, the alphabet in question would seem to be { `Aa, Bb, Cc, ..., :, !, ;, ...`}, that is, basically the characters on a computer keyboard. Let's denote this alphabet by 'E.' Elementary string theory tells us that though E^*, the set of all strings that can be built from E, is infinite, it's *countably* infinite, and that therefore there is a program P which enumerates E^*. (P, for example, can resort to lexicographic ordering.) From this it follows that the set of all stories is itself countably infinite.

However, though we concede there is good reason to think that the set of all stories is in some sense typographic, it needn't be countably infinite. Is the set $\mathcal{A}$ of all letter `A`s, countable? (Hofstadter (Hofstadter 1982) says "No.") If not, then simply imagine a story associated with every element within $\mathcal{A}$. For a parallel route to the same result, think of a story about π, a story about $\sqrt{2}$, indeed a story for every real number!

On the other hand, stories, in the real world, are often neither strings nor, more generally, typographic. After all, authors often think about, expand, refine, ... stories without considering anything typographic whatsoever. They may "watch" stories play out before their mind's eye, for example. In fact, it seems plausible to say that strings (and the like) can be used to *represent* stories, as opposed to saying that the relevant strings, strictly speaking, *are* stories.

function f from the set of all programs to Φ which, when given a candidate program P, yields an element of Φ for which P will fail. Put more formally (following (Dekker 1955), (Kugel 1986), (Post 1944)):

- Φ is Turing-undecidable if and only if $\exists f\ [f : \mathbf{TM} \rightarrow \Phi \wedge \forall m \in \mathbf{TM}\ \exists \phi \in \Phi\ (f(m) = \phi \wedge m \text{ cannot decide } \phi)]$

Put informally, a set Φ is productive if and only if it's not only classically undecidable, but also if any program proposed to decide Φ can be counter-exampled with some element of Φ. Evidence for the view that $\mathcal{I}$ is productive comes not only from the fact that even a descendant of BRUTUS$_1$ would seem to leave out some of $\mathcal{I}$, but from what the Autopoeisis team has experienced when attempting an outright definition of the locution 's is an interesting short-short story:' Every time a definition of this locution is ventured, someone comes up with a counter-example. (If it's proposed that all members of $\mathcal{I}$ must involve characters in some form of conflict, someone describes a monodrama wherein the protagonist is at utter, tedious peace. If it's proposed that all members of $\mathcal{I}$ must involve one or more key literary concepts (from a superset of those already mentioned: betrayal, unrequited love, etc.), someone describes an interesting story about characters who stand in a novel relationship. And so on.) So a key question arises: What about attempts to engineer creativity *without* trying to pre-represent the space of the artifacts desired?

6.2 A Non-Parameterized Evolutionary Approach

We all know that processing and representation are intimately linked.[25] So, given this general fact, how does one get the representation correct for creativity? If the representations used by BRUTUS are inadequate, what might work? What about creative processes in evolutionary computation? And what about marrying a new mode of representation to processing that is evolutionary in character, rather than (as in the case of BRUTUS) processing that is essentially theorem proving?

The first response to such questions is to observe that present systems of evolutionary computation would seem to squelch their capacity for creative processes because of their own impoverished representation schemes. Consider, for example, one of the standard processes of an evolutionary system: epigenesis — the process in an evolutionary system that translates a given genotype representation into the phenotype representation.[26] The present systems of evolutionary computation use parameterized computational structures to accomplish epigenesis: such systems use the genotype representation to encode levels of the parameters, and the phenotype representation becomes the direct output. The use of parameterized computational structures to formalize that which is within the reach of an evolutionary system

[25]For instance, although both decimal and roman numeral notations can represent numbers, the process of multiplying roman numerals is much more difficult than the process for multiplying decimals. Of the two notations, decimal notation is in general the better representation to enable mathematical processes (Marr 1982).

[26]We imagine some readers asking: "Mightn't 'morphogenesis' fit better?" We use 'epigenesis' in order to refer to the complete process of mapping from the genotype to the phenotype. However, morphogenesis does capture the essence of the process that is used in our (Noel's) work; frankly, we are smitten with the analogy. Choosing the atom features (say pixel for images, Hardy waves for sound) is similar to starting with a specialized cell, then forming the cell into some organization — morphogenesis.

works just as poorly as the methods behind BRUTUS. Pre-set parameterized computational structures are limited in their ability to map the levels of a finite number of parameters (even if the value of the parameters are infinite) onto a complete set of designs. Just as BRUTUS, for reasons discussed above, is tackling a space that probably can't be "mastered" via parameterized computational structures, so too present evolutionary systems face the same uphill battle. If we consider not $\mathcal{I}$, the space of interesting short-short stories, but the space of interesting *paintings*, then Hofstadter has probably given us the key paper: he has argued in (Hofstadter 1982) that when Knuth (1982) applied such a system to the design of fonts, he was bound to fail: the system could only cover a small portion of the set of all 'As,' for example. (The wildly imaginative As drawn in Hofstadter's paper are reason enough to read it; see Figure 3.) The use of parameterized computational structures, if you will, requires describing the quintessence of the hoped-for design before the (evolutionary) system can seek a solution.

Figure 3: Various Letter As

In our opinion, a system is creative only if it can somehow capture the quintessence of the space from which a particular design is is to come. On this view, creativity occurs outside of (at least most) current evolutionary systems. This is so because such systems, like the logic-based BRUTUS, are based on a pre-selected and bounded design space.

One of us (Noel), working with Sylvia Acchione-Noel, has created a system — MONA-LISA — that changes things: MONA-LISA uses an information-based representation that affects the resolution of the system (the ability to describe a wave form) but forces no feature-level dimension on the system. The representation is based on atomic or molecular representation, similar to the notions of atomic or molecular decomposition by Fourier analysis or wavelets (Meyer 1993). This new evolutionary system evolves patterns of pixels into any desired image. The use of a sub-feature representation requires the units to evolve simultaneously *en mass* to generate both the features and the configuration of an image. The evolved image is not constrained at the feature level and can encode a dog, a tree, a car, or, just as easily, a face. (If there is anything to the view that the set of all As makes a productive set, such a view in connection with the set of all faces is hardly implausible. And of course we believe that this view about As is quite plausible. We direct readers to (Hofstadter 1982) for the data and evidence. This paper includes an interesting collection of faces.) However, the resolution of the images is affected by the number of pixels or the atomic representation used in the system. For instance, the "portrait" of Abraham Lincoln shown in Figure 5 was evolved in a 25-by-25 pixel space which can only represent images with 12.5 lines of resolution or less.

Figure 4: Evolution of a Portrait of Lincoln

Traditional evolutionary programs use parameterized modeling to map between the genotype and the phenotype. The use of such models results in a direct mapping between the set of genes that comprise a parameter, and the level of the feature that the parameter models in the phenotype. There is no noise, no pheiotropy, no polygeny in the mapping. The

system can only create objects within the parameterized space, and all objects evolved are members of that space. For instance, one might model a face by making a model in which some genes selected types of eyes, noise, lips, ears, face shape, hair, etc., and other genes arrange the features within configurations of a normal face. If one were using such a system, the population of the first generation would all look like faces. One would select the faces that are most like the intended face and use them for breeding the next generation. The impact of this is that one must constantly compare faces with the intended face to make decisions. One face might have the eye shape and size right, while another might have the distance between the eyes correct.

In our technique, the desired or intended image is considered the signal, and all other images are considered noise. The elicitation of the image is done by a biased selection of the objects that generate the greatest recognition, or signal to noise ratio, in each generation. In keeping with our example, consider evolving a face image. The first generation is pure noise, or in other words, all possible images are equally likely. The task of the evolver is to select the images for breeding that have the greatest signal (most like the face to be evolved) and therefore the least noise. At first, the probability of any image looking like a face, any face, is extremely unlikely. Most images look like the snow on a TV tuned to a channel without a station. However, one can select images whose pixels might give the impression of something rounded in the middle, dark in the area of hair, or light where cheeks may be. Since the images selected to parent the next generation have more of the signal and consequently less noise, it will give rise to a population of images whose mean signal to noise ratio will be greater than the previous generation. Eventually the signal to noise ratio is strong enough to elicit the intended image.

In our evolutionary system one only sees the face that one seeks. The face is seen in different amounts of noise, from high to low. In high noise conditions, only the lowest spatial frequency in low contrast can be imaged. As the image evolves, the level and contrast of detail increases. The end image is much like seeing the intended face in a cloud in that there are no distinct features, but a holistic percept. While the quality of the image is at present limited in resolution (about 15 to 20 lines of resolution), the reader should be reminded that the system can evolve any image that the pixels can represent. One could just as easily select to evolve a horse, a ball, a tree, etc. In the traditional approach one is limited to a domain and would require a new model for each new type of object to be evolved.

Mona-Lisa is at present a two-dimensional system used by humans to create images, but it can be generalized to other domains. Humans were chosen to perform the judging and selecting.[27] Evolutionary systems are not necessarily, by themselves, creative. Creativity in evolution presumably occurs through the interaction of the objects and their environment under the forces of natural selection. However, as we have said, evolutionary systems *can* preclude creativity — by limiting and bounding the phenotype.[28] Mona-Lisa gives complete representational power to the user. It starts from scratch or from randomness, and it is truly

[27]Our intuition, for what it's worth, is that humans here provide a holistic evaluation function that mimics the forces of nature.

[28]Of course, even inspiration, insight, and phenomenal consciousness can preclude creativity, if one warps the example enough. But our point is really a practical warning: limiting and bounding the phenotype, *ceteris paribus*, can preclude creativity, so computational engineers beware!

general since the image's features are not pre-determined.[29]

Figure 5: Evolved Portrait of Lincoln

6.3 What Next?

Where do we intend to go from here? Both BRUTUS and MONA-LISA will live on, both to be gradually improved. In the case of BRUTUS, we are only at BRUTUS_1, and we will work sedulously as good engineers, clinging to a workday faith that some more robust system *can* capture all of $\mathcal{I}$. (When the workday ends, we will permit our suspicion that $\mathcal{I}$ is productive to have free rein.) In connection with MONA-LISA, we plan to

1. Attempt to substitute artificial agents (treated as "percept to action" functions (Russell & Norvig 1994)) for the role of the human in the image elicitation process.
2. Transfer the image elicitation approach to the domain of stories (so BRUTUS can get some help!).
3. Explore more carefully the complexity implications of our image elicitation work.[30]

7 Conclusion

Is $\text{Q1}_{\text{P}}^{\text{Z}}$ answered? Does our A3, bolstered by our engineering, really satisfy? Probably not — because not all will be convinced that creativity calls for "super"-information processing beyond what a zombie is allowed in scenarios like V2. However, tireless toil on BRUTUS and MONA-LISA may eventually reveal, for everyone to clearly see, that these projects are, at

[29]Here, for cognoscenti, are some more details on MONA-LISA: The DNA code is a vectorized string in which each gene represents one of the pixels in an associated image (usually a 25 x 25 pixel array). The level of a gene encodes the color of the pixel (usually 4 grays, or 8 colors). Fifty images are in each generation, of which the evolver selects ten to be the parents for the next generation. Reproduction is accomplished by selecting two parents at random and generating the offspring's DNA by randomly and uniformly selecting between the genes of the two parents at each allele site. Each population after the initial generation consists of the ten parents and forty offsprings allowing incest. MONA-LISA is a Boltzmann machine; as such its activation function is stochastic. Motivated readers may find it profitable to refer back to the brief account of neural nets given above, wherein the role of an activation function is discussed.

[30]We conclude with some brief remarks on point 3: As one might expect, an increase in the creative capacity of a system causes an increase in the system's complexity. Our evolutionary system creates representations with a new level of complexity over previous work in evolutionary computation. The increase in complexity is due to an increase in the cardinality of the relationships, increases in the level of emergent properties, and an increase in what Löfgren calls **interpretation and descriptive processes** (Löfgren 1974). The potential for complexity in a representation is determined by the relationships among the features. In the image elicitation system, the image is described at the atomic level. The low level description allows for an extremely large number of relationships as compared to systems that use a higher, feature-level representation. Image elicitation requires that features emerge along with the configuration, rather than serving to define the features with only the configuration evolving. As stated, our system promotes both polygenic and pleiotropic relationships between the genotype and the phenotype. Because of these relationships, the complexity of interpretation and description increases.

root, impossible, because they would have computation do what only "super"-computation can do. This would turn our engineering failure into philosophical gold.[31]

[31]For a complete treatment of super-computation and related matters, including literary creativity, see (Bringsjord & Zenzen 2001) and (Bringsjord 1998).

References

Ashcraft, M. (1994), *Human Memory and Cognition*, HarperCollins, New York, NY.

Averbach, E. & Coriell, A. S. (1961), 'Short-term memory in vision', *Bell System Technical Journal* **40**, 309–328.

Block, N. (1995), 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences* **18**, 227–247.

Boden, M. (1995), Could a robot be creative?—and would we know?, *in* K. Ford, C. Glymour & P. Hayes, eds, 'Android Epistemology', MIT Press, Cambridge, MA, pp. 51–72.

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

Bringsjord, S. (1997), 'Consciousness by the lights of logic and common sense', *Behavioral and Brain Sciences* **20.1**, 227–247.

Bringsjord, S. (1998), Philosophy and 'super' computation, *in* J. Moor & T. Bynam, eds, 'The Digital Phoenix: How Computers are Changing Philosophy', Blackwell, Oxford, UK, pp. 231–252.

Bringsjord, S. (1999), 'The zombie attack on the computational conception of mind', *Philosophy and Phenomenological Research* **59.1**, 41–69.

Bringsjord, S. & Ferrucci, D. (2000), *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Lawrence Erlbaum, Mahwah, NJ.

Bringsjord, S. & Zenzen, M. (2001), *SuperMinds: A Defense of Uncomputable Cognition*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Chalmers, D. (1996), *The Conscious Mind: In Search of a Fundmental Theory*, Oxford, Oxford, UK.

Cole, D. & Foelber, R. (1984), 'Contingent materialism', *Pacific Philosophical Quarterly* **65**(1), 74–85.

Cooper, L. & Shepard, R. (1973), Chronometric studies of the rotation of mental images, *in* W. Chase, ed., 'Visual Information Processing', Academic Press, New York, NY, pp. 135–142.

Davis, W. (1985), *The Serpent and the Rainbow*, Simon & Shuster, New York, NY.

Davis, W. (1988), *Passage of Darkness: The Ethnobiology of the Haitian Zombie*, University of North Caronlina Press, Chapel Hill, NC.

Dekker, J. C. E. (1955), 'Productive sets', *Transactions of the American Mathematical Society* **22**, 137–198.

Dennett, D. (1991), *Consciousness Explained*, Little, Brown, Boston, MA.

Dennett, D. (1995), 'The unimagined preposterousness of zombies', *Journal of Consciousness Studies* **2**(4), 322–326.

Dretske, F. (1996), 'Absent qualia', *Mind & Language* **11**(1), 78–85.

Earman, J. (1986), *A Primer on Determinism*, D. Reidel, Dordrecht, The Netherlands.

Fjelde, R. (1965), Foreward, *in* 'Four Major Plays — Ibsen', New American LIbrary, New York, NY, pp. ix–xxxv.

Flanagan, O. & Polger, T. (1995), 'Zombies and the function of consciousness', *Journal of Consciousness Studies* **2**(4), 313–321.

Harnad, S. (1995), 'Why and how we are not zombies', *Journal of Consciousness Studies* **1**, 164–167.

Hofstadter, D. (1982), 'Metafont, metamathematics, and metaphysics', *Visible Language* **14**(4), 309–338.

Knuth, D. (1982), 'The concept of a meta-font', *Visible Language* **14**(4), 3–27.

Kugel, P. (1986), 'Thinking may be more than computing', *Cognition* **18**, 128–149.

Löfgren, I. (1974), 'Complexity of descriptions of systems: A foundational study', *International Journal of General Systems* **3**, 197–214.

Marr, J. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Freeman, San Francisco, CA.

Meyer, Y. (1993), *Wavelets: Algorithms and applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Perry, J. (1979), 'The problem of the essential indexical', *Nous* **13**, 3–22.

Plum, F. & Posner, J. B. (1972), *The Diagnosis of Stupor and Coma*, F. A. Davis, Philadelphia, PA.

Pollock, J. (1995), *Cognitive Carpentry: A Blueprint for How to Build a Person*, MIT Press, Cambridge, MA.

Post, E. (1944), 'Recursively enumerable sets of positive integers and their decision problems', *Bulletin of the American Mathematical Society* **50**, 284–316.

Povinelli, D. (1997), What chimpanzees know about the mind, *in* 'Behavioral Diversity in Chimpanzees', Harvard University Press, Cambridge, MA, pp. 73–97.

Rosenthal, D. M. (1986), 'Two concepts of consciousness', *Philosohical Studies* **49**, 329–359.

Rosenthal, D. M. (1989), Thinking that one thinks, Technical Report 11, ZIF Report Zentrum für Interdisziplinäre Forschung, Bielefeld, Germany.

Rosenthal, D. M. (1990*a*), A theory of consciousness?, Technical Report 40, ZIF Report Zentrum für Interdisziplinäre Forschung, Bielefeld, Germany.

Rosenthal, D. M. (1990*b*), Why are verbally expressed thoughts conscious?, Technical Report 32, ZIF Report Zentrum für Interdisziplinäre Forschung, Bielefeld, Germany.

Rosenthal, D. M. (forthcoming), State consciousness and what it's like, *in* 'Title TBA', Clarendon Press, Oxford, UK.

Russell, S. & Norvig, P. (1994), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Saddle River, NJ.

Searle, J. (1992), *The Rediscovery of the Mind*, MIT Press, Cambridge, MA.

Shoemaker, S. (1975), 'Functionalism and qualia', *Philosophical Studies* **27**, 291–315.

Sperling, G. (1960), 'The information available in brief visual presentations', *Psychological Monographs* **74**, 48.

Thorndyke, P. W. (1977), Cognitive structures in comprehension and memory of narrative discourse, *in* 'Cognitive Psychology', Academic Press, New York, NY, pp. 121–152.