APA NEWSLETTER ON

Philosophy and Computers

Piotr Bołtuć, Editor | Fall 2009 | Volume 09, Number 1

FROM THE EDITOR

Piotr Bołtuć
University of Illinois at Springfield

A bloc of three important articles opens the current issue; they may be a little more formal than what some of our readers have come to expect. We are glad to feature an article by R. Turner, who asks what the meaning and the subject matter of a programming language is. He argues for a somewhat unconventional view that denotation of a programming language boils down to mathematical objects. No doubt Turner's elegant, formally sophisticated article will spark a lively debate on some central philosophical topics.

The two articles that follow—both by prominent computer scientists—engage in a dialogue with major philosophers of the past. G. Chaitin builds an argument pertaining to the complexity theory in constant dialogue with Leibniz. Chaitin claims that "Leibniz's ideas on complexity lead to a place where math seems to have no structure, none that we will ever be able to perceive." As a part of his interesting and lively argument Chaitin touches on the incompleteness theorem, the halting problem, and a number of other topics.

A. Sloman refers to Hume so as to carefully carve a theoretical niche for his main thesis that "an organism (human or non-human) or machine may have…motive whose existence is merely a product of the operation of a motive-generating mechanism" resulting from evolution, human design, or some contingency. Sloman claims that Hume leaves some room for "reflexes" as an avenue for naturalistic motivations (though they are marginal in humans, and so in Hume's theory) and that what Sloman calls architecture based motivation is such a reflex mechanism.

*

I was glad to present the last few issues of this Newsletter at the closing session of the North American Computers and Philosophy conference in Bloomington Indiana recently.[1] I focused on three main topic-areas highlighted in this Newsletter; I also emphasized the role of editor-reviewed, or one-way-blind-reviewed, publications in the life of the profession.

The main topic-area to which I think this Newsletter has had something to contribute lately falls between philosophy of mind and machine consciousness. We started with discussion of the explanatory gap (G. Harman, Y. Nagasawa, et al.) which then gained further momentum when it blended, somewhat, with the discussion of the role of LIDA for machine consciousness (I. Aleksander and those mentioned below). I am glad that in the current issue we have another part of this debate where S. Franklin, B. Baars, and U. Ramamurthy respond to their critics while P. Haikonen responds back.

The second topic area pertains to L. Floridi's view on information ethics as ethics tout court (Floridi, T. Baynum, et al.). In the current issue we have a follow up on this discussion in K. Herold's critique of J. Barker's position and Barker's spirited response.

The third major topic-area is ontology of artificial objects. We started this conversation with the question of the ontological status of artifacts (Baker, Thomasson, et al.). In the current issue we focus on the ontological status of web-based objects in thought provoking papers by D. L. Anderson and R. Arp. Anderson tackles directly the issue of the ontological status of objects that function in virtual environments, such as The World of Warcraft and Second Life. The paper builds, impressively, on a rare combination of competencies in analytical ontology and familiarity with virtual environments. Arp tries to bridge the gap between the understandings of ontology in informatics and philosophy. Philosophers will find an informative presentation of formal ontologies (primarily domain ontologies) used in IT while information scientists may find the standard arguments for ontological realism, the view which seems underrepresented in their discipline.

*

The second part of this Newsletter is by far less analytical; we have one phenomenological, two feminist, and one Marxist paper, if those stickers mean anything anymore. We are pleased to continue the conversation about aesthetics and computers so aptly started by Lopes and Matravers, which we have undertaken in conjunction with ASA. In the current issue G. Tavinor focuses on the various senses of interactivity in art and whether videogames are any special in this regard.

The two articles on online education and gender issues were a part of an APA session organized by this committee at the Pacific APA meeting (A. White moderator). M. A. Crouch focuses on specificities of gender in online education. It seems that if there is still gender discrimination in the classroom, there is less of it online; on the other hand, there may be ramifications of getting any education for people who work and take primary responsibility for their families, predominantly women. H. E. Baber discusses the reasons why women do not blog as much as men. Baber argues that women still have reasons to be more guarded than men in their impromptu online contributions since they may be interpreted through a different, largely less favorable lens. A future study could perhaps show why more women than men are present on certain social networking sites, such as Twitter. We also have a review-article of a new book by C. Fuchs on how social networking sites may be used for surveillance and other controversial purposes. The author uses critical theory, developed by the Frankfurt School, to analyze the positive and negative aspects of social networking in a broader socio-economic context.

Last, but by far not least, we close this issue with A. Sloman's long document about teaching AI and philosophy in

schools. The motivation behind this project is the perception, shared by many teachers and students, that AI is just a helpful skill with not enough deep theoretical grounding to merit the treatment as a discipline at the university level. Sloman's goal is to meet this challenge head on.

*

As always, I want to close by thanking my Dean, as well as my Department Chair, at the University of Illinois at Springfield for making it possible for me to devote more attention to this Newsletter than I would have been able to do otherwise. Let me end with a special note. Now when the APA site is relatively in order[2] I would want to guide the Readers towards this Newsletter's history. The older issues, I think especially those edited by Jon Dorbolo,[3] will remain an excellent source of information about the history of philosophy and computing and are still very much worth browsing through.

Endnotes

1. I want to thank T. Beavers for his impromptu invitation during my visit there.
2. http://www.apaonline.org/publications/newsletters/computers.aspx
3. Jon assembled an excellent team of editors: W. Uzgalis, R. Causey, L. Hinman (as the Internet Resources Editor) and R. Barnette (as the Teaching in Cyberspace Editor). See http://www.apaonline.org/publications/newsletters/v97n1_Computers_01.aspx

From the Chair

Michael Byron
Kent State University

The Pacific Division held its 2009 meeting in April in Vancouver, BC. Last fall, the Committee voted to award the Barwise Prize to Terry Bynum of Southern Connecticut State. Unfortunately, the session had to be canceled due to travel issues. We expect to reschedule this award at the 2010 Central Division Meeting to be held next February in Chicago.

For 2009-10 the Committee welcomes two new members. David L. Anderson is in the philosophy department at Illinois State and directs the Mind Project there (http://bit.ly/DGqAH). Susan V.H. Castro (http://bit.ly/Stmia) completed her Ph.D. recently at UCLA. We are always glad to have new members.

The next divisional meeting will be the Eastern Division Meeting, to be held in December in New York City. At that meeting we will be presenting the Barwise Prize to Luciano Floridi. Beyond that, the Committee looks forward to another productive year in 2009-10.

New and Noteworthy: A Central APA Invitation

We want to invite you to two sessions at the 2010 APA Central Division meeting in Chicago.

1. North American Computing and Philosophy Conference, (NA-CAP) Thursday, February 18, 7:30-10:30 p.m.:

Machine Consciousness

Chair:

Marvin Croy

Papers:

Ricardo Sanz (Technical University of Madrid): "The Need for a Mind in Control Systems Engeneering"

Piotr Boltuc (University of Illinois at Springfield): "Non-Reductive Machine Consciousness?"

Matthias Scheutz (Indiana University–Bloomington): "Architectural Steps Towards Self-Aware Robots"

Commentators:

Thomas Polger (University of Cincinnati)

John Barker (University of Illinois at Springfield)

2. Special Session Organized by the APA Committee on Philosophy and Computers, Saturday, February 20 (morning)

Machines, Intentionality, Ethics and Cognition

Chair:

Peter Boltuc (University of Illinois at Springfield)

Participants:

David L. Anderson (Illinois State University–Bloomington): "Why Intentional Machines Must be Moral Agents (or at least Moral Patients)"

Keith Miller W. (University of Illinois at Springfield): "Truth in Advertising, or Disrespecting Robot Autonomy"

Svetoslav Braynov (University of Illinois at Springfield): "Can you Trust a Robot?"

Thomas Polger (University of Cincinnati): "Distributed Computation and Extended Cognition"

Closing comments:

Ricardo Sanz (Technical University of Madrid)

Articles

Featured Article

The Meaning of Programming Languages

Raymond Turner
University of Essex

Abstract

A folklore view has it that programming languages get their semantic interpretations layer by layer, one language getting its interpretation in the next, until the bedrock of physical reality (physical machines) provides the final and actual mechanism of semantic interpretation. We argue, based upon the normative requirements of any semantic account, that this is a false picture. We further argue that, in any adequate semantic theory of a programming language, the denotations of its constructs must be taken to be mathematical objects.

1. The Limitations of Grammar

Programming languages form part of the bedrock on which computer science is built. Their design and implementation is a core aspect of the subject, and they are the host for the day to day activity of program and software construction. Consequently, a proper conceptual analysis of the nature of these languages will form a significant part of the philosophy

of computer science. While a great many of the conceptual questions that surround them focus on their design, and their use in problem solving and software construction, here we shall concentrate on some of the issues that center upon their definition, and, in particular, on their semantic import.

In practice, programming languages get their semantic interpretation using a mixture of methods. Often, a top level natural language account guides the construction of a compiler that translates the language into the language of the implementing machine. This may or may not be direct, i.e., the interpretation may pass through several layers and several different languages, and associated compilers. It may even be cushioned by the presence of an intermediate abstract machine, but ultimately this process ends in the machine instructions of a physical machine. But how do these techniques fix the semantic content of the language? Part of our objective is to examine this question. But first we must set the scene.

Generally, a programming language, as a language, is given via a formal grammar of some sort. This spells out the legal strings of symbols of the language. For example, the following provides the syntax for a simple imperative language in a standard recursive notation, where P stands for programs, E for arithmetic expressions, and B for Boolean expressions.

P ::= x:= E | **skip**| P; P | **if** B **then** P **else** P | **while** B **do** P

E ::= x |0|1| E+E | E*E

B ::= x | true | false |E < E| ¬B|B ∨ B

The expressions (E) are constructed from 0 and 1 by addition and multiplication. The Boolean expressions (B) are constructed from true, false, the ordering relation (<) on numbers and negation and conjunction. The actual programs of the language (P) are built from a simple assignment statement (x:= E) via sequencing (P; Q), conditional programs (**if** B **then** P **else** Q) and while loops (**while** B **do** P).[1] For example, according to this grammar, the following is a grammatically legal program.

x:=0; (1)
y:=1;
while x < n **do** (x:=x+1; y:= x*y)

But by itself, the grammar does not tell us what this program does or what it is supposed to do. If you have grasped its semantic import, it is because you already have some understanding of the intended computational impact of its constructs. The point is, to construct or understand this program, or any program for that matter, one needs to know more than the syntax of its host language; one must possess some semantic information about the language.

2. The Normative Nature of Semantics

Any such semantics must provide an account of the intended meaning of the constructs. This must be sufficient to guide a compiler writer in implementing the language and facilitate arbitration when disputes arise over the implementation of a construct: it must enable a distinction to be drawn between the correct and incorrect implementation of a construct. It must also support a distinction between correct and incorrect programs—not just syntactically, but in the sense of meeting their intended specifications. For instance, a semantic account must determine that program (1) with input n, computes the factorial function. Likewise, it must determine that

x := 0; y := 1;
if n = 0 **then** y := x + 1;
while x > n **do** (x := x–1; y := (x*y)+y) (2) does not.

Any semantic account must act as a normative guide to the language: it must enable a distinction to be drawn between a correct and incorrect use of a language construct. This seems to apply even to natural language.[2]

> Suppose the expression "green" means green. It follows immediately that the expression "green" applies correctly only to these things (the green ones) and not to those (the non-greens). The fact that the expression means something implies, that there is a whole set of normative truths about my behavior with that expression: namely, that my use of it is correct in application to certain objects and not in application to others. ...The normativity of meaning turns out to be, in other words, simply a new name for the familiar fact that, regardless of whether one thinks of meaning in truth-theoretic or assertion-theoretic terms, meaningful expressions possess conditions of correct use. (On the one construal, correctness consists in true use, on the other, in warranted use.) Kripke's insight was to realize that this observation may be converted into a condition of adequacy on theories of the determination of meaning: any proposed candidate for the property in virtue of which an expression has meaning, must be such as to ground the "normativity" of meaning—it ought to be possible to read off from any alleged meaning constituting property of a word, what is the correct use of that word. [2]

This normativity constraint seems not to be a controversial one for semantics in general. However, in the case of programming languages, it has some clear, yet significant implications for any proposed semantic account.

Presumably, language designers have some semantic intentions about the computational impact of their language constructs, and one way such intentions might be articulated is via an informal natural language account of the various constructs, where these descriptions most often take the form of a reference manual for the language. And for real languages these often run to hundreds of pages, e.g., the specification of the Java Language is almost 600 pages.

What do these language specifications look like? Normally they are provided in terms of the impact of the language constructs upon an underlying machine. For our simple language we require a machine with an underlying state whose role is to store numerical values in locations, i.e., a state of the machine might look like the following.

[3 4 7 . . .]
[x y z]

where the visual display of the numerals indicates their numerical content. The full recursive language is then interpreted via its impact upon this machine. But before we embark on any further elaboration of this, we have to face a preliminary question.

Is this to be taken as a physical or abstract machine? Some authors ([4], [3], [5]), suggest that programming language constructs are ambiguous, i.e., they have two meanings, one provided by an abstract machine and one provided by physical one. For example, according to the latter, the assignment statement

x := 10 (3)

is given its interpretation by its impact upon a physical device, i.e., where its meaning is given as

place 10 in location x,

x refers to the physical location. What are the consequences of this? Presumably, that somehow the meaning is given and fixed by the physical machine. What else could it mean? Consequently,

although the intentions of the machine designer may have been used to guide the construction of the physical machine, they are no longer definitive. If when run, the instruction x := 10 sticks 20 in location x, then so be it; this is what it is taken to mean. The intentions of the designer are superseded by the actual impact of the instructions on the machine. But this has an important consequence, namely, there is no notion of malfunction. There is no alternative court of appeal. If, during the running of a machine instruction, the machine switches on and off, this is to be taken as part of the meaning of the instruction. But is this a coherent perspective? The rule following considerations ([27], [2]), and Kripke suggest not.

> Actual machines can malfunction: through melting wires or slipping gears they may give the wrong answer. How is it determined when a malfunction occurs? By reference to the program of the machine, as intended by its designer, not simply by reference to the machine itself. Depending on the intent of the designer, any particular phenomenon may or may not count as a machine malfunction. A programmer with suitable intentions might even have intended to make use of the fact that wires melt or gears slip, so that a machine that is malfunctioning for me is behaving perfectly for him. Whether a machine ever malfunctions and, if so, when, is not a property of the machine itself as a physical object but is well defined only in terms of its program, stipulated by its designer. Given the program, once again, the physical object is superfluous for the purpose of determining what function is meant. [11] page 34

The notion of malfunction must be measured against a stable account that reflects the designer's intentions. And this cannot be supplied by the physical machine. Of course, we could impose some restrictions on the device to ensure that it behaves appropriately. We might, for instance, suppose that there are physical mechanisms that enable us to perform an Update and a Lookup on the state of the machine, and these must satisfy the following requirement.

> Suppose in the state s, the machine is updated by inserting v in location x. If in the resulting state, the value in location x is then looked up, then the value v will be returned. If the value in location y (where y is different to x) is looked up, then the value of y in the original state s is returned.

We can put this a little more precisely. We shall use the phrase E evaluates to v to indicate that the expression E reduces to the value v at the end of the computation.

> Lookup(Update(s, x, v), x) evaluates to v
>
> If Lookup(s, y) evaluates to w then Lookup(Update(s, x, v), y) evaluates to w – where x and y are distinct.

So that if Lookup(Update(s, x,10)) evaluates to 20, then the physical machine has malfunctioned. But this is a definition of an abstract machine. Indeed, as Kripke observed, given the abstract machine, from the semantic perspective, the physical one is superfluous. It is the abstract machine that serves as the basis for semantics, and as a guide to the construction of the physical one.

So what does this say about the so-called physical interpretation of assignment given by (3)? Clearly, it cannot function as a definitional account of the construct. So what is its relationship to the abstract normative one? Only that the normative meaning has physical implications, i.e., against the background of its normative meaning, the physical implications may be used to predict the behavior of the physical machine. In this simple case, it is the physical machine that is under scrutiny. (3) does not have two definitional readings. Indeed, why is this case not parallel to any mathematical notion and its application to the physical world? We do not feel compelled to say that the notion of a right angled triangle has two meanings: the definitional one and the one that results from its consequences when applied to the physical world. The latter does not yield a second meaning.

This is one step in our argument to the effect that any semantic theory must be an abstract one: at its base there must be an abstract machine.

3. An Informal Semantics

With this much established, we may provide some account of the semantics of our language: relative to it, we provide an evaluation mechanism for the programs of our simple language.

1. If the evaluation of E in the state s returns the value v, then the evaluation of x := E in a state s, returns the state that is the same as s except that the value v replaces the current value in location x.
2. The evaluation of **skip** in a state s, returns s.
3. If the evaluation of B in s returns true and the evaluation of P in s returns s′, then the evaluation of if B then P else Q in s, evaluates to s′. If on the other hand, the evaluation of B in s returns false and the evaluation of Q in s returns s′, then the evaluation of **if** B **then** P **else** Q in s, returns s′.
4. If the evaluation of P in s yields the state s′ and the evaluation of Q in s′ returns the state s″, then the evaluation of P; Q in s, returns the state s″.
5. If the evaluation of B in s returns true, the evaluation of P in s returns s′, and the evaluation of **while** B **do** P in s′ yields s″, then the evaluation of **while** B **do** P in s, returns s″. If the evaluation of B in s returns false, then the evaluation of **while** B **do** P in s, returns s.

There are several assumptions built into 1-5 that need to be made explicit. First, observe that the evaluation of Boolean and numerical expressions is assumed not to change the state. This is a property of the language. Indeed, it is an essential property for the coherence of the semantics given by 1-5. We shall say more about such properties later. Secondly, the semantics allows for the possibility that programs may not terminate; in such cases the premises of the informal rules may not be true.

Such accounts provide our first foothold on grasping the semantic impact of the constructs of the language. But does the informality of such an account undermine its normative role? Surely normative stipulation must be exact enough to decide what is right and what is wrong. Do natural language accounts enable the articulation of a semantic theory that is simultaneously precise and transparent? Does the act of removing all possible ambiguities render the semantics unreadable and opaque? While this is not a serious issue for the present toy language, it appears to be so for commercial ones. For example, the original natural language description of Java was seriously flawed [17]. Of course, we might try to provide a more precise account by employing a programming language in which to write the semantic description. But this would just push the semantic problem onto a new language.[3]

4. A More Formal Account

There are many formal approaches ([1], [15], [23], [21], [24], [22]) but their advocates are united in the belief that natural language is not a suitable vehicle for expressing combinatorial

notions. To illustrate matters, and to serve as a vehicle for our investigation, we shall provide a slightly more formal version of our informal account. We shall write

$$< P, s > \Downarrow s'$$

to indicate that evaluating P in state s terminates in s′ [6]. Indeed, we can almost read off the formal rules from our informal account.

1. The memory update command has the following rule of evaluation

$$\frac{< E, s > \Downarrow v}{< x := E, s > \Downarrow \text{Update}(s, x, v)}$$

The premise guarantees that the expression E in state s reduces to value v. The conclusion then guarantees that program x := E will update the state s with value v in the location x.

2. The skip operation is given by the following transition rule.

$$\frac{}{< \textbf{skip}, s > \Downarrow s}$$

3. Sequencing is governed by the following pair of rules.

$$\frac{< P, s > \Downarrow s' \qquad < Q, s' > \Downarrow s''}{< P; Q, s > \Downarrow s''}$$

The premise of the first rule insists that the program P in state s returns the state s′; and the second that Q in state s′ yields state s″. The conclusion then guarantees that the program C; D in state s will return the state s″.

4. The conditional is governed by the following two rules that cover the true and false cases for the evaluation of the Boolean expression.

$$\frac{< B, s > \Downarrow \text{true} \qquad < P, s > \Downarrow s'}{< \textbf{If } B \textbf{ do } P \textbf{ else } Q, s > \Downarrow s'}$$

$$\frac{< B, s > \Downarrow \text{false} \qquad < Q, s > \Downarrow s''}{< \textbf{If } B \textbf{ do } P \textbf{ else } Q, s > \Downarrow s''}$$

5. Finally, we provide the rules of the while command. The first rule deals with the case where the Boolean is true and the second where it evaluates to false. Note that the premise in the first assumes that the while loop terminates in the state derived from evaluating P.

$$\frac{< B, s > \Downarrow \text{true} \qquad < P, s > \Downarrow s' \qquad < \textbf{while } B \textbf{ do } P, s' > \Downarrow s''}{< \textbf{while } B \textbf{ do } P, s > \Downarrow s''}$$

$$\frac{< B, s > \Downarrow \text{false}}{< \textbf{while } B \textbf{ do } P, s > \Downarrow s}$$

The semantics is structural, because the meaning of a program is defined by the meaning of its components. This form of (big step) Structural Operational Semantics (SOS) was introduced by Gordon Plotkin [16]. It provides a clear guide to the implementor without imposing how exactly the latter is to proceed on a given concrete machine.

Indeed, the semantics can be looked at as an axiomatic theory of operations: the language enables the construction of complex operations from simple ones, and the rules provide their content. In particular, they tell us what the termination conditions for the constructs are. Of course, in order to constitute a useful theory, it requires some development. In particular, we may introduce a notion of equivalence.

$$P \bullet Q \times \text{œ} s_1 \text{A } \text{œ} s_2 \text{A } < P, s_1 > \Downarrow s_2 \leftrightarrow < Q, s_1 > \Downarrow s_2$$

i.e., two programs are behaviorally equivalent if they behave identically, i.e., started in the same state, they terminate in the same state. This provides us with a notion of partial equality. Given this we may show that the following rules may be established

(i) **while** B **do** P • **if** B **then** (P; **while** B **do** P) **else skip**

(ii) **if** true **then** P **else** Q • P

(iii) **if** false **then** P **else** Q • Q

And though not a deep and exciting theory, this is beginning to look much like any other mathematical theory. Indeed, we seemed to have arrived at the conclusion that any purported semantic theory, i.e., one that meets our normativity requirements, must be mathematical in nature. It would seem to follow that programming languages are definitionally, mathematical objects, i.e., their very definition as semantic objects makes them so. We shall refer to this as our central claim. Unfortunately, there are some possible objections. We shall consider these in the next two sections.

5. Informal Mathematics

One concerns the move from informal to formal semantics. The practicing programmer (or even compiler writer) might well claim that formal accounts are unnecessary. Indeed, our move from the informal account to the formal one was not justified by the present language, but by the vagaries that might creep in with real, large, and complex languages. However, so it may be argued, while one may have to take great care in formulating matters, and mistakes may be made, even for the most syntactically complex languages, there is nothing in principle that blocks the development of an informal normative account.

Does this mean that our central claim fails? No. We may accept this criticism but argue that even informal accounts are mathematical. More exactly, we might be persuaded that the move to a formal semantics is unnecessary, but still claim that the informal semantics is mathematical. To see why such account must be taken to be mathematical, consider again the problem of its coherence. To establish that latter, we need to make explicit some of the rules for the evaluation of expressions.

1. To evaluate a variable, in a state s, lookup its value. And return the same state.
2. If in state s, E evaluates to (v, s′) and E′ in state s′ evaluates to (v′, s″) then E + E′ evaluates to (v + v′, s″) and E * E′ evaluates to (v * v′, s″).

To show that this coheres with 1-5 for the evaluation of programs, we need to show that expression evaluation does not change the state. For this, we argue by induction on the structure of expressions. The case of variables is clear. Here there is no change to the state, just a simple lookup. Moreover, if two expressions do not change the state, then neither does their sum or product. The same may be argued for Boolean expressions. This is consistent with the evaluation of Boolean expressions in the evaluation of conditional programs. Consequently, the whole semantics fits together.

At some level such checking is not really optional; it is an integral part of the activity of specifying the language. But the important point is that such arguments are mathematical; just because they are expressed in English does not mean that they are not so. Indeed, they involve a form of structural induction, and it would be a strange notion of mathematics that ruled them out. Furthermore, most of actual mathematics is conducted at this level of informality. Few areas of mathematics are practiced with the rigor of formal logic. Indeed, even our informal semantic account provides, albeit informal, an axiomatic theory of operations. Moreover, many other theories have started out as informal axiomatic theories, and are only later made more precise when the appropriate formal language

is invented. Geometry is an obvious example. It was and still is mathematical, and its entities are mathematical ones.

So the informality of the semantics does not obviously undermine the claim that normative accounts of programming languages will render them mathematical objects.

6. Are Operational Accounts Mathematical?

However, some would deny this. They would do so, not by drawing a distinction between our formal and informal accounts, but by insisting that neither is mathematical. In ([21], [14]) such a criticism is aimed at Landin's operational approach [12].

> We can apparently get quite a long way expounding the properties of a language with purely syntactic rules and transformations.....But we must remember that when working like this all we are doing is manipulating symbols—we have no idea at all of what we are talking about. To solve any real problem, we must give some semantic interpretation. [21]

> The motivation for the Scott-Strachey approach to semantics stems from a recognition of the fact that programs and the objects they manipulate are in some sense concrete realisations or implementations of abstract mathematical objects. [14]

Apparently, operational accounts do not provide an interpretation into the abstract mathematical world of numbers and sets. Likewise, our rules of evaluation do not yield a mathematical account. In the end, we are still left with uninterpreted systems of rules. And a formal system, with no intended interpretation, is still an uninterpreted formal system, i.e., the rules themselves need to be interpreted. Moreover, any such interpretation relies on a further interpretation that relies on more rules, etc. Consequently, so the argument goes, we still do not have an account where this regress is blocked by reference to abstract mathematical objects.[4] If this is so, while we have provided a system of rules that constitute a normative account of the language, this system does not constitute a mathematical theory.

Apparently, for a semantic account to be mathematical, it must be based upon mathematical objects. For example, one such would have the semantics associate, with each program, a mathematical function from states to states, i.e., the semantic function C would have the form

$$C : \text{Program} \Rightarrow (\text{State} \Rightarrow \text{State})$$

where State is a set and State ⇒ State represents some class of set theoretic functions ([26], [1],[10]). Underlying this demand is the view that set theory is taken to be more than just a formal theory. In particular, the language of set theory is taken to refer beyond syntax to the abstract world of sets, an objective world, independent of language and our knowledge of it. Hence, it is taken to block any such regress of languages. Its language refers to an abstract pre-existing world of sets [13]. In particular, the axioms of set theory are taken to be justified by the picture of the cumulative hierarchy.

In general, genuine mathematical systems are taken to point beyond the syntax of the formal system to an abstract mathematical world. This is a realist view of mathematical structures ([13], [20]). In contrast, our fledgling theory of operations, which is a theory of operations that is determined by the programs of our simple language, is taken not to be a theory that refers to an abstract world. Indeed, if it were, it would seem to follow that every programming language gives rise to such a theory. And does this not lead to the implausible view that programming languages are not designed but discovered? Well, not quite.

There might be an abstract theory of operations that is given independently of these languages. A theory from which programming language designers select constructs. This seems to have been the view of Strachey [22]. Indeed, put this way, it is unclear why this is less plausible than the case of set theory. We certainly require some positive metaphysical reason to put set theory in a different class to theories of operations. Gödel provides a possible one.[5]

> Despite their remoteness from sense experience, we do have something like a perception also of the objects of set-theory, as is seen from the fact that the axioms force themselves upon us as being true. I don't see any reason why we should have less confidence in this kind of perception, i.e., in mathematical intuition than in sense perception, which induces us to build up physical theories and to expect that future sense perceptions will agree with them, and moreover, to believe that a question not decidable now has meaning and may be decided in the future. [9]

But, even if we accept something like it, there is still a puzzle about the difference between a theory of operations and a theory of sets; even under such an interpretation, it seems hard to see how the difference could be made out. It is certainly not clear that Gödel would have supported such a distinction. Referring to Turing's analysis of mechanical computability, he writes:

> The greatest improvement was made possible through the precise definition of the concept of finite procedure, which plays a decisive role in these results. There are several different ways of arriving at such definition, which, however, all lead to the same concept. The most satisfactory way, in my opinion, is that of reducing the concept of finite procedure to that of a machine with a finite number of parts, as has been done by the British mathematician Turing. [8]

Consequently, one assumes that he would have assigned the notion of finite procedure a similar metaphysical status to sets. In Wang's words, Gödel saw the problem of defining computability as:

> an excellent example of a concept which did not appear sharp to us but has become so as a result of a careful reflection. [25]

It would seem that any theory of operations that achieved definitive status would have the same metaphysical status as sets. Indeed, if one finds the concept of pre-existing notions of operations implausible, it is hard to see how one can defend the view of a pre-existing world of sets [24]. It would seem that the two theories stand or fall together: either they are both mathematical theories or both are not. But since set theory is a paradigm case of a mathematical theory, to deny it mathematical status would be to deny almost anything mathematical status.

But the story does not end here. Even though the case of mathematical status for these toy languages seems plausible, whether this can be maintained for real languages is less clear.

Acknowledgements

Rosana Turner provided detailed and very insightful comments on various drafts of the paper.

References

[1] Allison, L. 1990. A Practical Introduction to Denotational Semantics. Cambridge: Cambridge Computer Science Texts.

[2] Boghossian, P.A. 1989. The rule-following considerations. Mind 89.

[3] Colburn, T. 2004. Methodology of Computer Science. The Blackwell Guide to the Philosophy of Computing and Information, edited by Luciano Floridi. 318-26. Malden, MA: Blackwell.

[4] Colburn, T.R. 2000. Philosophy and computer science. Explorations in Philosophy Series. New York: M.E. Sharpe.

[5] Fetzer, J.H. 1988. Program verification: the very idea. Communications of the ACM 31(9): 1048-63.

[6] Fernandez. M. 2007. Programming Languages and Operational Semantics: An Introduction. Oxford.

[7] Gödel, K. 1934. Undecidable diophantine propositions. In Collected Works III. 164-75.

[8] Gödel, K. 1934. Some basic theorems on the foundations of mathematics and their implications. In Collected Works III. 304-23.

[9] Gödel, K. 1983. What is Cantor's continuum problem? Reprinted in Benacerraf and Putnam's collection Philosophy of Mathematics, 2nd ed., Cambridge University Press.

[10] http://en.wikibooks.org/wiki/Haskell/Denotational_semantics

[11] Kripke, S. 1982. Wittgenstein on Rules and Private Language. Harvard University Press.

[12] Landin, P.J. 1964. The mechanical evaluation of expressions. The Computer Journal 6(4): 308-20; doi:10.1093/comjnl/6.4.308.

[13] Maddy, P. 1992. Realism in Mathematics. Oxford: Oxford University Press.

[14] McGettrick, A.D. 1980. The Definition of Programming Languages. Cambridge University Press New York, NY: Cambridge University Press.

[15] Milne, R. and Strachey, C. 1977. A Theory of Programming Language Semantics. New York, NY: Halsted Press.

[16] Plotkin, G. 2004. A structural approach to operational semantics. J. Log. Algebr. Program 60-61: 17-139.

[17] Pugh, W. 2000. The Java, memory model is fatally flawed. Concurrency: Practice and Experience 12(6): 445-55.

[18] Rapaport, W.J. 1999. Implementation is semantic interpretation. Monist 82: 109-30.

[19] Rapaport, W.J. 2005. Implementation is semantic interpretation: Further Thoughts. Journal of Experimental and Theoretical Artificial Intelligence 17(4): 385-417.

[20] Shapiro, S. 2004. Philosophy of Mathematics: Structure and Ontology. Oxford.

[21] Stoy, J. 1977. The Scott-Strachey Approach to Programming Language Semantics. MIT Press.

[22] Strachey, C. 1966. Towards a formal semantics. In Formal Language Description Languages for Computer Programming. 198-220. North Holland.

[23] Tennent, R.D. 1991. Semantics of Programming Languages. London: Prentice-Hall International.

[24] Turner, R. 2007. Understanding programming languages. Minds and Machines 17(2): 129-33.

[25] Wang, H. 1974. From Mathematics to Philosophy. London: Routledge & Kegan Paul.

[26] http://en.wikipedia.org/wiki/Denotational_semantics

[27] Wittgenstein, L. 1953. Philosophical Investigations. Oxford: Blackwell Publishing.

Endnotes

1. We shall use parenthesis to disambiguate.
2. Presumably, in natural language, the semantics does not play a definitional role. It merely codifies what we take to be correct and incorrect use.
3. [18], [19] offer a detailed discussion of the nature of implementation as semantic interpretation.
4. Moreover, at this point in the argument, we have no other obvious way out of the regress since we have given up the bedrock of physical reality as a mechanism to fix the meaning.
5. Maddy [13] defends a version of Gödel's view.

Leibniz, Complexity, and Incompleteness[1]

Gregory Chaitin
IBM Research

Let me start with Hermann Weyl, who was a fine mathematician and mathematical physicist. He wrote books on quantum mechanics and general relativity. He also wrote two books on philosophy: The Open World: Three Lectures on the Metaphysical Implications of Science (1932), a small book with three lectures that Weyl gave at Yale University in New Haven, and Philosophy of Mathematics and Natural Science, published by Princeton University Press in 1949, an expanded version of a book he originally published in German.

In these two books Weyl emphasizes the importance for the philosophy of science of an idea that Leibniz had about complexity, a very fundamental idea. The question is what is a law of nature, what does it mean to say that nature follows laws? Here is how Weyl explains Leibniz's idea in The Open World, pp. 40-41: The concept of a law becomes vacuous if arbitrarily complicated laws are permitted, for then there is always a law. In other words, given any set of experimental data, there is always a complicated ad hoc law. That is valueless; simplicity is an intrinsic part of the concept of a law of nature.

What did Leibniz actually say about complexity? Well, I have been able to find three or perhaps four places where Leibniz says something important about complexity. Let me run through them before I return to Weyl and Popper and more modern developments.

First of all, Leibniz refers to complexity in Sections V and VI of his 1686 Discours de métaphysique, notes he wrote when his attempt to improve the pumps removing water from the silver mines in the Harz mountains was interrupted by a snow storm. These notes were not published until more than a century after Leibniz's death. In fact, most of Leibniz's best ideas were expressed in letters to the leading European intellectuals of his time, or were found many years after Leibniz's death in his private papers. You must remember that at that time there were not many scientific journals. Instead, European intellectuals were joined in what was referred to as the Republic of Letters. Indeed, publishing could be risky. Leibniz sent a summary of the Discours de métaphysique to the philosophe Arnauld, himself a Jansenist fugitive from Louis XIV, who was so horrified at the possible heretical implications that Leibniz never sent the Discours to anyone else. Also, the title of the Discours was supplied by the editor who found it among Leibniz's papers, not by Leibniz.

I should add that Leibniz's papers were preserved by chance, because most of them dealt with affairs of state. When Leibniz died, his patron, the Duke of Hanover, by then the King of England, ordered that they be preserved, sealed, in the Hanover royal archives, not given to Leibniz's relatives. Furthermore, Leibniz produced no definitive summary of his views. His ideas

are always in a constant state of development, and he flies like a butterfly from subject to subject, throwing out fundamental ideas, but rarely, except in the case of the calculus, pausing to develop them.

In Section V of the Discours, Leibniz states that God has created the best of all possible worlds, in that all the richness and diversity that we observe in the universe is the product of a simple, elegant, beautiful set of ideas. God simultaneously maximizes the richness of the world, and minimizes the complexity of the laws which determine this world. In modern terminology, the world is understandable, comprehensible, science is possible. You see, the Discours was written in 1686, the year before Leibniz's nemesis Newton published his Principia, when medieval theology and modern science, then called mechanical philosophy, still coexisted. At that time the question of why science is possible was still a serious one. Modern science was still young and had not yet obliterated all opposition.

The deeper idea, the one that so impressed Weyl, is in Section VI of the Discours. There Leibniz considers "experimental data" obtained by scattering spots of ink on a piece of paper by shaking a quill pen. Consider the finite set of data points thus obtained, and let us ask what it means to say that they obey a law of nature. Well, says Leibniz, that cannot just mean that there is a mathematical equation passing through that set of points, because there is always such an equation! The set of points obey a law only if there is a **simple** equation passing through them, not if the equation is "fort composée" = very complex, because then there is always an equation.

Another place where Leibniz refers to complexity is in Section 7 of his Principles of Nature and Grace (1714), where he asks why is there something rather than nothing, why is the world non-empty, because "nothing is simpler and easier than something!" In modern terms, where does the complexity in the world come from? In Leibniz's view, from God; in modern terminology, from the choice of the laws of nature and the initial conditions that determine the world. Here I should mention a remarkable contemporary development: Max Tegmark's amazing idea that the ensemble of all possible laws, all possible universes, is simpler than picking any individual universe. In other words, the multiverse is more fundamental than the question of the laws of our particular universe, which merely happens to be our postal address in the multiverse of all possible worlds! To illustrate this idea, the set of **all** positive integers 1, 2, 3, ... is very simple, even though **particular** positive integers such as 9859436643312312 can be arbitrarily complex.

A third place where Leibniz refers to complexity is in Sections 33-35 of his Monadology (1714), where he discusses what it means to provide a mathematical proof. He observes that to prove a complicated statement we break it up into simpler statements, until we reach statements that are so simple that they are self-evident and don't need to be proved. In other words, a proof reduces something complicated to a consequence of simpler statements, with an infinite regress avoided by stopping when our analysis reduces things to a consequence of principles that are so simple that no proof is required.

There may be yet another interesting remark by Leibniz on complexity, but I have not been able to discover the original source and verify this. It seems that Leibniz was once asked why he had avoided crushing a spider, whereupon he replied that it was a shame to destroy such an intricate mechanism. If we take "intricate" to be a synonym for "complex," then this perhaps shows that Leibniz appreciated that biological organisms are extremely complex.

These are the four most interesting texts by Leibniz on complexity that I've discovered. As my friend Stephen Wolfram has remarked, the vast Leibniz Nachlass may well conceal other treasures, because editors publish only what they can understand. This happens only when an age has independently developed an idea to the point that they can appreciate its value plus the fact that Leibniz captured the essential concept.

Having told you about what I think are the most interesting observations that Leibniz makes about simplicity and complexity, let me get back to Weyl and Popper. Weyl observes that this crucial idea of complexity, the fundamental role of which has been identified by Leibniz, is unfortunately very hard to pin down. How can we measure the complexity of an equation? Well, roughly speaking, by its size, but that is highly time-dependent, as mathematical notation changes over the years and it is highly arbitrary which mathematical functions one takes as given, as primitive operations. Should one accept Bessel functions, for instance, as part of standard mathematical notation?

This train of thought is finally taken up by Karl Popper in his book The Logic of Scientific Discovery (1959), which was also originally published in German, and which has an entire chapter on simplicity, Chapter VII. In that chapter Popper reviews Weyl's remarks, and adds that if Weyl cannot provide a stable definition of complexity, then this must be very hard to do.

At this point these ideas temporarily disappear from the scene, only to be taken up again, to reappear, metamorphized, in a field that I call **algorithmic information theory** (AIT). AIT provides, I believe, an answer to the question of how to give a precise definition of the complexity of a law. It does this by changing the context. Instead of considering the experimental data to be points, and a law to be an equation, AIT makes everything digital, everything becomes 0s and 1s. In AIT, a law of nature is a piece of software, a computer algorithm, and instead of trying to measure the complexity of a law via the size of an equation, we now consider the size of programs, the number of bits in the software that implements our theory:

Law: Equation → Software,

Complexity: Size of equation → Size of program, Bits of software.

The following diagram illustrates the central idea of AIT, which is a very simple toy model of the scientific enterprise:

Theory (01100...11) → **COMPUTER** → Experimental Data (110...0).

In this model, both the theory and the data are finite strings of bits. A theory is software for explaining the data, and in the AIT model this means the software produces or calculates the data exactly, without any mistakes. In other words, in our model a scientific theory is a program whose output is the data, self-contained software, without any input.

And what becomes of Leibniz's fundamental observation about the meaning of "law?" Before there was always a complicated equation that passes through the data points. Now there is always a theory with the same number of bits as the data it explains, because the software can always contain the data it is trying to calculate as a constant, thus avoiding any calculation. Here we do not have a law; there is no real theory. Data follows a law, can be understood, only if the program for calculating it is much smaller than the data it explains.

In other words, understanding is compression, comprehension is compression, a scientific theory unifies many seemingly disparate phenomena and shows that they reflect a common underlying mechanism.

To repeat, we consider a computer program to be a theory for its output, that is the essential idea, and both theory and

output are finite strings of bits whose size can be compared. And the best theory is the smallest program that produces that data, that precise output. That's our version of what some people call Occam's razor. This approach enables us to proceed mathematically, to define complexity precisely and to prove things about it. And once you start down this road, the first thing you discover is that most finite strings of bits are lawless, algorithmically irreducible, algorithmically random, because there is no theory substantially smaller than the data itself. In other words, the smallest program that produces that output has about the same size as the output. The second thing you discover is that you can never be sure you have the best theory.

Before I discuss this, perhaps I should mention that AIT was originally proposed, independently, by three people, Ray Solomonoff, A. N. Kolmogorov, and myself, in the 1960s. But the original theory was not quite right. A decade later, in the mid 1970s, what I believe to be the definitive version of the theory emerged, this time independently due to me and to Leonid Levin, although Levin did not get the definition of relative complexity precisely right. I will say more about the 1970s version of AIT, which employs what I call "self-delimiting programs," later, when I discuss the halting probability Ω.

But for now, let me get back to the question of proving that you have the best theory, that you have the smallest program that produces the output it does. Is this easy to do? It turns out this is extremely difficult to do, and this provides a new complexity-based view of incompleteness that is very different from the classical incompleteness results of Gödel (1931) and Turing (1936). Let me show you why.

First of all, I'll call a program "elegant" if it's the best theory for its output, if it is the smallest program in your programming language that produces the output it does. We fix the programming language under discussion, and we consider the problem of using a formal axiomatic theory, a mathematical theory with a finite number of axioms written in an artificial formal language and employing the rules of mathematical logic, to prove that individual programs are elegant. Let's show that this is hard to do by considering the following program **P**:

P produces the output of the first provably
elegant program that is larger than **P**.

In other words, **P** systematically searches through the tree of all possible proofs in the formal theory until it finds a proof that a program **Q**, that is larger than **P**, is elegant, then **P** runs this program **Q** and produces the same output that **Q** does. But this is impossible, because **P** is too small to produce that output! **P** cannot produce the same output as a provably elegant program **Q** that is larger than **P**, not by the definition of elegant, not if we assume that all provably elegant programs are in fact actually elegant. Hence, if our formal theory only proves that elegant programs are elegant, then it can only prove that finitely many individual programs are elegant.

This is a rather different way to get incompleteness, not at all like Gödel's "This statement is unprovable" or Turing's observation that no formal theory can enable you to always solve individual instances of the halting problem. It's different because it involves complexity. It shows that the world of mathematical ideas is infinitely complex, while our formal theories necessarily have finite complexity. Indeed, just proving that individual programs are elegant requires infinite complexity. And what precisely do I mean by the complexity of a formal mathematical theory? Well, if you take a close look at the paradoxical program **P** above, whose size gives an upper bound on what can be proved, that upper bound is essentially just the size in bits of a program for running through the tree of all possible proofs using mathematical logic to produce all the theorems, all the consequences of our axioms. In other words, in AIT the complexity of a math theory is just the size of the smallest program for generating all the theorems of the theory.

And what we just proved is that if a program **Q** is more complicated than your theory **T**, **T** can't enable you to prove that **Q** is elegant. In other words, it takes an **N**-bit theory to prove that an **N**-bit program is elegant. The Platonic world of mathematical ideas is infinitely complex, but what we can know is only a finite part of this infinite complexity, depending on the complexity of our theories.

Let's now compare math with biology. Biology deals with very complicated systems. There are no simple equations for your spouse, or for a human society. But math is even more complicated than biology. The human genome consists of 3×10^9 bases, which is 6×10^9 bits, which is large, but which is only finite. Math, however, is infinitely complicated, provably so.

An even more dramatic illustration of these ideas is provided by the halting probability Ω, which is defined to be the probability that a program generated by coin tossing eventually halts. In other words, each **K**-bit program that halts contributes 1 over 2^K to the halting probability Ω. To show that Ω is a well-defined probability between zero and one it is essential to use the 1970s version of AIT with self-delimiting programs. With the 1960s version of AIT, the halting probability cannot be defined, because the sum of the relevant probabilities diverges, which is one of the reasons it was necessary to change AIT.

Anyway, Ω is a kind of DNA for pure math, because it tells you the answer to every individual instance of the halting problem. Furthermore, if you write Ω's numerical value out in binary, in base-two, what you get is an infinite string of irreducible mathematical facts:

$$\Omega = .11011...$$

Each of these bits, each bit of Ω, has to be a 0 or a 1, but it's so delicately balanced, that we will never know. More precisely, it takes an **N**-bit theory to be able to determine **N** bits of Ω.

Employing Leibnizian terminology, we can restate this as follows: The bits of Ω are mathematical facts that refute the principle of sufficient reason, because there is no reason they have the values they do, no reason simpler than themselves. The bits of Ω are in the Platonic world of ideas and therefore **necessary** truths, but they look very much like **contingent** truths, like accidents. And that's the surprising place where Leibniz's ideas on complexity lead, to a place where math seems to have no structure, none that we will ever be able to perceive. How would Leibniz react to this?

First of all, I think that he would instantly be able to understand everything. He knew all about 0s and 1s, and had even proposed that the Duke of Hanover cast a silver medal in honor of base-two arithmetic, in honor of the fact that everything can be represented by 0s and 1s. Several designs for this medal were found among Leibniz's papers, but they were never cast, until Stephen Wolfram took one and had it made in silver and gave it to me as a sixtieth birthday present. And Leibniz also understood very well the idea of a formal theory as one in which we can mechanically deduce all the consequences. In fact, the calculus was just one case of this. Christian Huygens, who taught Leibniz mathematics in Paris, hated the calculus, because it was mechanical and automatically gave answers, merely with formal manipulations, without any understanding of what the formulas meant. But that was precisely the idea, and how Leibniz's version of the calculus differed from Newton's. Leibniz invented a notation which led you automatically, mechanically, to the answer, just by following certain formal rules.

And the idea of computing by machine was certainly not foreign to Leibniz. He was elected to the London Royal Society, before the priority dispute with Newton soured everything, on the basis of his design for a machine to multiply. (Pascal's original calculating machine could only add.)

So I do not think that Leibniz would have been shocked; I think that he would have liked Ω and its paradoxical properties. Leibniz was open to all systèmes du monde, he found good in every philosophy, ancient, scholastic, mechanical, Kabbalah, alchemy, Chinese, Catholic, Protestant. He delighted in showing that apparently contradictory philosophical systems were, in fact, compatible. This was at the heart of his effort to reunify Catholicism and Protestantism. And I believe it explains the fantastic character of his Monadology, which, complicated as it was, showed that certain apparently contradictory ideas were, in fact, not totally irreconcilable.

I think we need ideas to inspire us. And one way to do this is to pick heroes who exemplify the best that mankind can produce. We could do much worse than pick Leibniz as one of these exemplifying heroes.[2]

Endnotes

1. Lecture given Friday, June 6, 2008, at the University of Rome "Tor Vergata," in a meeting on "Causality, Meaningful Complexity, and Knowledge Construction." I thank Professor Arturo Carsetti for inviting me to give this talk.
2. For more on such themes, please see Chaitin, Meta Maths, Atlantic Books, London, 2006, or the collection of my philosophical papers, Chaitin, Thinking about Gödel and Turing (Singapore: World Scientific, 2007).

Architecture-Based Motivation vs. Reward-Based Motivation

Aaron Sloman
University of Birmingham

Introduction

"Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them." David Hume, A Treatise of Human Nature (2.3.3.4), 1739-1740 (http://www.class.uidaho.edu/mickelsen/ToC/hume%20treatise%20ToC.htm)

Whatever Hume may have meant by this, and whatever various commentators may have taken him to mean, I claim that there is at least one interpretation in which this statement is obviously true, namely: no matter what factual information an animal or machine A contains, and no matter what competences A has regarding abilities to reason, to plan, to predict, or to explain, A will not actually **do** anything unless it has, in addition, some sort of control mechanism that selects among the many alternative processes that A's information and competences can support.

In short: control mechanisms are required in addition to factual information and reasoning mechanisms if A is to do anything. This paper is about what forms of control are required. I assume that in at least some cases there are motives, and the control arises out of selection of a motive for action. That raises the question where motives come from. My answer is that they can be generated and selected in different ways, but one way is not itself motivated: it merely involves the operation of mechanisms in the architecture of A that generate motives and select some of them for action. The view I wish to oppose is that all motives must somehow serve the interests of A, or be rewarding for A. This view is widely held and is based on a lack of imagination about possible designs for working system. I summarize it as the assumption that all motivation must be reward-based. In contrast, I claim that at least some motivation may be architecture-based, in the sense explained below.

Instead of talking about "passions," I shall use the less emotive terms "motivation" and "motive." A motive in this context is a specification of something to be done or achieved (which could include preventing or avoiding some state of affairs, or maintaining a state or process). The words "motivation" and "motivational" can be used to describe the states, processes, and mechanisms concerned with production of motives, their control and management, and the effects of motives in initiating and controlling internal and external behaviors. So Hume's claim, as interpreted here, is that no collection of beliefs and reasoning capabilities can generate behavior on its own: motivation is also required.

This view of Hume's claim is expressed well in the Stanford Encyclopedia of Philosophy entry on motivation, though without explicit reference to Hume:

> The belief that an antibiotic will cure a specific infection may move an individual to take the antibiotic, if she also believes that she has the infection, and if she either desires to be cured or judges that she ought to treat the infection for her own good. All on its own, however, an empirical belief like this one appears to carry with it no particular motivational impact; a person can judge that an antibiotic will most effectively cure a specific infection without being moved one way or another. (http://plato.stanford.edu/entries/moral-motivation)

That raises the question: Where do motives come from and why are some possible motives (e.g., going for lunch) selected and others (e.g., going for a walk, or starting a campaign for election to parliament) not selected?

If Hume had known about reflexes, he might have treated them as an alternative mode of initiation of behavior to motivation (or passions). There may be some who regard a knee-jerk reflex as involving a kind of motivation produced by tapping a sensitive part of the knee. That would not be a common usage. I think it is more helpful to regard such physical reflexes as different from motives, and therefore as exceptions to Hume's claim. I shall try to show that something like "internal reflexes" in an information-processing system can be part of the explanation of creation and adoption of motives. In particular, adopting the "design-based approach to the study of mind" yields a wider variety of possible explanations of how minds work than is typically considered in philosophy or psychology, and paradoxically even in AI/Robotics, where such an approach ought to be more influential.

This proposal opposes a view that all motives are selected on the basis of the costs and benefits of achieving them, which we can loosely characterize as the claim that all motivation is "reward-based."

In the history of philosophy and psychology there have been many theories of motivation, and distinctions between different sorts of motivation, for example, motivations related to biological needs, motivations somehow acquired through cultural influences, motivations related to achieving or maximizing some reward (e.g., food, admiration in others, going to heaven), or avoiding or minimizing some punishment (often labelled positive and negative reward or reinforcement), motivations that are means to some other end, and motivations that are desired for their own sake, motivations related to intellectual or other achievements, and so on. Many theorists assume that motivation must be linked to rewards or utility. One

version of this (a form of hedonism) is the assumption that all actions are done for ultimately selfish reasons.

I shall try to explain why there is an alternative kind of motivation, architecture-based motivation, which is not included even in this rather broad characterization of types of motivation on Wikipedia:

> Motivation is the set of reasons that determines one to engage in a particular behavior. The term is generally used for human motivation but, theoretically, it can be used to describe the causes for animal behavior as well. This article refers to human motivation. According to various theories, motivation may be rooted in the basic need to minimize physical pain and maximize pleasure, or it may include specific needs such as eating and resting, or a desired object, hobby, goal, state of being, ideal, or it may be attributed to less-apparent reasons such as altruism, morality, or avoiding mortality. (http://en.wikipedia.org/wiki/Motivation)

Philosophers who write about motivation tend to have rather different concerns such as whether there is a necessary connection between deciding what one morally ought to do and being motivated to do it. For more on this see the afore-mentioned entry in the Stanford Encyclopedia of Philosophy.

Motivation is also a topic of great concern in management theory and management practice, where motivation of workers comes from outside them, e.g., in the form of reward mechanisms (providing money, status, recognition, etc.) sometimes in other forms, e.g., inspiration, exhortation, social pressures. I shall not discuss any of those ideas.

In psychology and even in AI, all these concerns can arise, though I am here only discussing questions about the mechanisms that underlie processes within an organism or machine that select things to aim for and which initiate and control the behaviors that result. This includes mechanisms that produce goals and desires, mechanisms that identify and resolve conflicts between different goals or desires, mechanisms that select means to achieving goals or desires.

Achieving a desired goal G could be done in different ways, e.g.,

- select and use an available plan for doing things of type G
- use a planning mechanism to create a plan to achieve G and follow it
- detect and follow a gradient that appears to lead to achieving G (e.g., if G is being on high ground to avoid a rising tide, walk uphill while you can)

There is much more to be said about the forms different motives can have, and the various ways in which their status can change, e.g., when a motive has been generated but not yet selected, when it has been selected, but not yet scheduled, or when there is not yet any clear plan or strategy as to how to achieve it, or whether action has or has not been initiated, whether any conflict with other motives, or unexpected obstacle has been detected, etc.

For a characterization of some of the largely unnoticed complexity of motives see http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#16.

L.P. Beaudoin, A. Sloman, A study of motive processing and attention, Prospects for Artificial Intelligence, IOS Press, 1993 (further developed in Luc Beaudoin's Ph.D. thesis).

Where do motives come from?

It is often assumed that motivation, i.e., an organism's or machine's, selection, maintenance, or pursuance of some state of affairs, the motive's content, must be related to the organism or machine having information (e.g., a belief or expectation) that achievement of the motive will bring some rewards or benefit, sometimes referred to as "utility." This could be reduction of some disadvantage or disutility, e.g., a decrease in danger or pain.

Extreme versions of this assumption are found in philosophical theories that all agents are ultimately selfish, since they can only be motivated to do things that reward themselves, even if that is a case of feeling good about helping someone else.

More generally, the assumption is that selection of a motive among possible motives must be based on some kind of prediction about the consequences of achieving or preventing whatever state of affairs is specified in that motive. This document challenges that claim by demonstrating that it is possible for an organism or machine to have, and to act on, motives for which there is no such prediction.

My claim

My claim is that an organism (human or non-human) or machine may have something as a motive whose existence is merely a product of the operation of a motive-generating mechanism—which itself may be a product of evolution, or something produced by a designer, or something that resulted from a learning or developmental process, or, in some cases, may be produced by some pathology. Where the mechanism comes from and what its benefits are are irrelevant to its being a motivational mechanism: all that matters is that it should generate motives, and thereby be capable of influencing selection of behaviors.

In other words, it is possible for there to be reflex mechanisms whose effect is to produce new motives, and in simple cases to initiate behaviors controlled by such motives. I shall present a very simple architecture illustrating this possibility below, though for any actual organism, or intelligent robot, a more complex architecture will be required, for reasons given later.

Where the reflex mechanisms come from is a separate question: they may be produced by a robot designer or by biological evolution, or by a learning process, or even by some pathology (e.g., mechanisms producing addictions) but what the origin of such a mechanism is, is a separate question from what it does, how it does it, and what the consequences are.

I am not denying that some motives are concerned with producing benefits for the agent. It may even be the case (which I doubt) that most motives generated in humans and other animals are selected because of their benefit for the individual. For now, I am merely claiming that something different can occur and does occur, as follows:

> Not all the mechanisms for generating motives in a particular organism O, and not all the motives produced in O have to be related to any reward or positive or negative reinforcement for O.
>
> What makes them motives is how they work: what effects they have, or, in more complex cases, what effects they tend to have even though they are suppressed (e.g., since competing, incompatible, motives can exist in O).

Learning and motivation

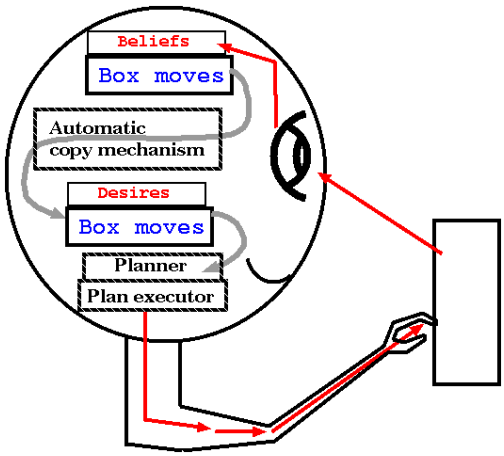
Many researchers in AI and other disciplines (though not all) assume that learning must be related to reward in some way, e.g., through positive or negative reinforcement.

I think that is false: some forms of learning occur simply because the opportunity to learn arises and the information-processing architecture produced by biological evolution simply reacts to many opportunities to learn, or to do things that could produce learning because the mechanisms that achieve that have proved their worth in previous generations, without the animals concerned knowing that they are using those mechanisms nor why they are using them.

Architecture-based motivation

Consider a very simple design for an organism or machine (Figure 1). It has a perceptual system that forms descriptions of a process occurring in the environment. Those descriptions are copied/stored in a database of "current beliefs" about what is happening in the world or has recently happened.

Figure 1. A simple design for an organism or machine.



At regular intervals another mechanism selects one of the beliefs about processes occurring recently and copies its content (perhaps with some minor modification or removal of some detail, such as direction of motion) to form the content of a new motive in a database of "desires." The desires may be removed after a time.

At regular intervals an intention-forming mechanism selects one of the desires to act as a goal for a planning mechanism that works out which actions could make the desire come true, selects a plan, then initiates plan execution.

This system will automatically generate motives to produce actions that repeat or continue changes that it has recently perceived, possibly with slight modifications, and it will adjust its behaviors so as to execute a plan for fulfilling the latest selected motive.

Why is a planning mechanism required instead of a much simpler reflex action mechanism that does not require motives to be formulated and planning to occur?

A reflex mechanism would be fine if evolution had detected all the situations that can arise and if it had produced a mechanism that is able to trigger the fine details of the actions in all such situations. In general that is impossible, so instead of a process automatically triggering behavior it can trigger the formation of some goal to be achieved, and then a secondary process can work out how to achieve it in the light of the then current situation.

For such a system to work there is NO need for the motives selected or the actions performed to produce any reward. We have goals generated and acted on without any reward being required for the system to work. Moreover, a side effect of such processes might be that the system observes what happens when these actions are performed in varying circumstances, and thereby learns things about how the environment works. That can be a side effect without being an explicit goal.

A designer could put such a mechanism into a robot as a way of producing such learning without that being the robot's goal. Likewise, biological evolution could have selected changes that lead to such mechanisms existing in some organisms because they produce useful learning, without any of the individual animals knowing that they have such mechanisms nor how they were selected or how they operate.

More complex variations

There is no need for the motive generating mechanism to be so simple. Some motives triggered by perceiving a physical process could involve systematic variations on the theme of the process, e.g., undoing its effects, reversing the process, preventing the process from terminating, joining in and contributing to an ongoing process, or repeating the process, but with some object or action or instrument replaced. A mechanism that could generate such variations would accelerate learning about how things work in the environment, if the effects of various actions are recorded or generalized or compared with previous records, generalizations, and predictions.

The motives generated will certainly need to change with the age and sophistication of the learner.

Some of the motive-generating mechanisms could be less directly triggered by particular perceived episodes and more influenced by the previous history of the individual, taking account not only of physical events but also social phenomena, e.g., discovering what peers seem to approve of, or choose to do. The motives generated by inferring motives of others could vary according to stage of development. For example, early motives might mainly be copies of inferred motives of others, then as the child develops the ability to distinguish safe from unsafe experiments, the motives triggered by discovering motives of others could include various generalizations or modifications, e.g., generalizing some motive to a wider class of situations, or restricting it to a narrower class, or even generating motives to oppose the perceived motives of others (e.g., parents!).

Moreover, some of the processes triggered instead of producing external actions could produce internal changes to the architecture or its mechanisms. Those changes could include production of new motive generators, or motive comparators, or motive generator generators, etc.

For more on this idea see chapter 6 and chapter 10 of The Computer Revolution in Philosophy (1978).

Mechanisms required

In humans it seems that architecture-based motivation plays a role at various levels of cognitive development, and is manifested in early play and exploration, and in intellectual curiosity later on, e.g., in connection with things like mathematics or chess, and various forms of competitiveness.

Such learning would depend on other mechanisms monitoring the results of behavior generated by architecture-based motivational mechanisms and looking for both new generalizations, new conjectured explanations of those generalizations, and new evidence that old theories or old conceptual systems are flawed—and require debugging.

Such learning processes would require additional complex mechanisms, including mechanisms concerned with construction and use of powerful forms of representation and mechanisms for producing substantive (i.e., non-definitional) ontology extension.

For more on additional mechanisms required see http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang

Evolution of minds and languages. What evolved first and develops first in children: Languages for communicating, or languages for thinking (Generalised Languages: GLs)

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#prague09

Ontologies for baby animals and robots. From "baby stuff" to the world of adult science: Developmental AI from a Kantian viewpoint.

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#toddlers

A New Approach to Philosophy of Mathematics: Design a young explorer, able to discover "toddler theorems" (Or: "The Naive Mathematics Manifesto").

The mechanisms constructing architecture-based motivational sub-systems could sometimes go wrong, accounting for some pathologies, e.g., obsessions, addictions, etc. But at present that is merely conjecture.

Conclusion

If all this is correct, then humans, like many other organisms, may have many motives that exist not because having them benefits the individual but because ancestors with the mechanisms that produce those motives in those situations happened to produce more descendants than conspecifics without those mechanisms did. Some social insect species in which workers act as "slaves" serving the needs of larvae and the queen appear to be examples. In those cases it may be the case that

Some motivational mechanisms "reward" the genomes that specify them, not the individuals that have them.

Similarly, some forms of learning may occur because animals that have certain learning mechanisms had ancestors who produced more offspring than rivals that lacked those learning mechanisms. This could be the case without the learning mechanism specifically benefiting the individual. In fact, the learning mechanism may lead to parents adopting suicidal behaviors in order to divert predators from their children.

It follows that any AI and cognitive science research based on the assumption that learning is produced ONLY by mechanisms that maximize expected utility for the individual organism or robot is likely to miss out on important forms of learning. Perhaps the most important forms.

One reason for this is that typically individuals that have opportunities to learn do not know enough to be able to even begin to assess the long-term utility of what they are doing. So they have to rely on what evolution has learnt (or a designer in the case of robots) and, at a later stage, on what the culture has learnt. What evolution or a culture has learnt may, of course, not be appropriate in new circumstances!

This discussion note does not prove that evolution produced organisms that make use of architecture-based motivation in which at least some motives are produced and acted on without any reward mechanism being required. But it illustrates the possibility, thereby challenging the assumption that ALL motivation must arise out of expected rewards.

Similar arguments about how suitably designed reflex mechanisms may react to perceived processes and states of affairs by modifying internal information stores could show that at least some forms of learning use mechanisms that are not concerned with rewards, with positive or negative reinforcement, or with utility maximization (or maximization of expected utility). My conjecture is that the most important forms of learning in advanced intelligent systems (e.g., some aspects of language learning in human children) are architecture-based, not reward based. But that requires further investigation.

The ideas presented here are very relevant to projects like CogX, which aim to investigate designs for robots that "self-understand" and "self-extend," since it demonstrates at least the possibility that some forms of self-extension may not be reward-driven, but architecture-driven.

Various forms of architecture-based motivation seem to be required for the development of precursors of mathematical competences described here: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#toddlers.

Some of what is called "curiosity-driven" behavior probably needs to be re-described as "architecture-based" or "architecture-driven."

This is one of a series of notes explaining how learning about underlying mechanisms can alter our views about the "logical topography" of a range of phenomena, suggesting that our current conceptual schemes (Gilbert Ryle's "logical geography") can be revised and improved, at least for the purposes of science, technology, education, and maybe even for everyday conversation, as explained in http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html.

Note

Marvin Minsky wrote quite a lot about goals and how they are formed in The Emotion Machine. It seems to me that the above is consistent with what he wrote, though I may have misinterpreted him.

Something like the ideas presented here were taken for granted when I wrote The Computer Revolution in Philosophy in 1978. However, at that time I underestimated the importance of spelling out assumptions and conjectures in much greater detail.

Acknowledgements

I wish to thank Veronica Arriola Rios and Damien Duff for helpful comments on an earlier, less clear draft.

DISCUSSION 1: ON ROBOT CONSCIOUSNESS

Robots Need Conscious Perception: A Reply to Aleksander and Haikonen

Stan Franklin
University of Memphis

Bernard J. Baars
The Neuroscience Institute, San Diego

Uma Ramamurthy
St. Jude Children's Research Hospital

In response to our article entitled "A Phenomenally Conscious Robot?" Igor Aleksander and Pentti Haikonen were kind enough to write responses (Aleksander 2009; Haikonen 2009).

Aleksander prefers to start with phenomenal consciousness from the start of the modeling process rather than adding on to a functional model. We have no preference in that respect. Global Workspace Theory (GWT) is based on a vast empirical literature with phenomenal experiences as a major testable ingredient (Baars 2002). LIDA began "life" as a functional model. We're happy to start either way, as long as the result is a working model that also reflects both phenomenal and third-person evidence about consciousness. Let's aim at modeling phenomenal consciousness.

Our hypothesis is that both GWT as fleshed out in the LIDA model and a coherent perceptual field will prove to be necessary conditions for phenomenal consciousness. This doesn't assert that "phenomenal states are added to functional structures…" Such functional structures can well be a necessary attribute of phenomenal consciousness without phenomenal states being "add-ons."

The essence of each of Aleksander's five axioms for phenomenally conscious states seems to lie in the notion of "feeling." Aleksander's term "feelings" is nothing but phenomenal consciousness. This is the famous Hard Problem of consciousness, but Aleksander does not give us an answer to it. It is not clear how GWT-LIDA is expected to solve the Hard Problem if no one else, it is claimed, can do that.

It is only this requirement that prevents a LIDA controlled, functionally conscious software agent from satisfying all five axioms. Put another way, we believe that given time and resources, producing such a functionally conscious software agent or robot, based on the LIDA architecture, that satisfies all five axioms except, perhaps, for the "feeling" requirement, would be a relatively straightforward project. The question of how to determine whether or not such a software agent or robot would have "feelings" would remain, or it might just fade, as has happened with the definition of "the essence of life" and other putatively impossible scientific questions.

Aleksander further asserts that "a neural substrate appears to be necessary to satisfy several aspects of the axioms." He goes on to assert that "…the vividness of phenomenal experience is helped by creating neural state machines with states that use a large number of neurons as state variables." "Helped," yes; but "necessary," is not at all clear. It seems at least plausible to us that such vividness could be achieved in a software agent or robot using a large number of virtual state variables.

Aleksander claims that a sufficiently complex neural network is needed to achieve the perceptual detail and resolution of phenomenal consciousness. He asserts that this "…may be difficult to design within a functionalist framework." We believe that a functional model, like LIDA, can be designed to achieve any necessary level of perceptual resolution.

Aleksander next discusses "program branching," asserting that:

> A non-neural functionalist representation would be more like frames in movies on film, which branch only through some very smart recognition of some features in stored "coherent perceptual fields" rather than a system whose state structure directly reflects the dynamic, branching experience.

"Program branching" assumes an old-fashioned AI symbolic agent. A LIDA controlled agent would perceive partially through a slipnet whose recognition occurs via passing of activation from primitive feature detectors, much like neural processing, and not like symbolic AI (Franklin 2005; Mitchell 1993). There are no "stored images" or "stored 'coherent perceptual fields'."

In his conclusion Aleksander claims that "It is very difficult to start with a model based in classical cognitive science, the mode of expression of which is algorithmic and implies virtual systems determined through the intention of a programmer." Assuming that the model described in the previous sentence is intended to be the LIDA model controlling an autonomous agent, the assertion significantly misrepresents LIDA. A LIDA controlled agent is provided by a programmer with sensing capabilities (sensors, primitive feature detectors, etc.), action capabilities (effectors), motivators (feeling/emotions), and a basic cognitive cycle, including several modes of learning, with which to answer the continual, primary question for every autonomous agent, "What shall I do next?" Evolutionary processes provide biological agents, such as humans, with these exact same elements for the exact same purpose. A LIDA based agent in a complex, dynamically changing environment must go through a developmental period as would a human child, and would continue learning thereafter. Again, the LIDA model seems to have been confused with classical symbolic AI models.

In his response Haikonen writes:

> What is functional consciousness? Franklin, Baars, and Ramamurthy answer: "An agent is said to be functionally consciousness (sic) if its control structure implements the Global Workspace Theory and the LIDA Cognitive Cycle." However, this is not a proof. This is a definition and as such conveniently eliminates the need to study if there were anything in this implementation that could even remotely qualify as and resemble functional consciousness. This kind of study might be difficult, because in nature there is no such thing as functional consciousness.

There are no proofs in science in the sense of evidence so strong as to not be susceptible to challenge. Such proofs belong only in mathematics. Every scientific "fact" is continually open to challenge. Correct mathematical proofs are not.

There is a sizable and growing body of evidence from cognitive science and neuroscience that human minds (their control structures) implement the essential elements of Global Workspace Theory (Baars 2002; Gaillard et al. 2009) and the LIDA Cognitive Cycle (Canolty et al. 2006; Jensen & Colgin 2007; Massimini et al. 2005; Uchida, Kepecs, & Mainen 2006; van Berkum 2006; Willis & Todorov 2006). This satisfies our definition of functional consciousness (Franklin 2003).

Haikonen goes on to assert that:

> On the other hand, Merker (2005) has proposed that phenomenal consciousness produces a stable perceptual world by distinguishing real motion from the apparent motion produced by the movement of the sensors. Franklin reads this proposition backwards and concludes that phenomenal consciousness can be produced by the production of stable perceptual world.

Franklin concluded no such thing. Haikonen begins his response by correctly asserting that the FBR paper "propose[s] that providing a functionally conscious robot with stable coherent perceptual world might be a step towards a phenomenally conscious machine." Note his own phrase "might be a step toward."

Accusing FBR of faulty logic, Haikonen claims that "…stable perception cannot be a cause for phenomenal consciousness. Merker's original proposition and Franklin's conclusion must be suspected." There was no faulty logic since the stated conclusion was never drawn. Also, it seems possible that a stable perceptual world might be part of a sufficient set of conditions for phenomenal consciousness, without being necessary. In other words, removing the stable perceptual world condition from the sufficient set might render it no longer sufficient, without the removed condition being necessary for phenomenal consciousness.

References

Aleksander, Igor. 2009. Essential phenomenology for conscious machines: a note on Franklin, Baars and Ramamurthy: "A Phenomenally Conscious Robot." APA Newsletter on Philosophy and Computers 08:2.

Baars, Bernard J. 2002. The conscious access hypothesis: origins and recent evidence. Trends in Cognitive Science 6: 47-52.

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., et al. 2006. High gamma power is phase-locked to theta oscillations in human neocortex. Science 313: 1626-28.

Franklin, S. 2003. IDA: a conscious artifact? Journal of Consciousness Studies 10: 47-66.

Franklin, S. 2005. A "consciousness" based architecture for a functioning mind. In Visions of Mind, edited by Darryl N. Davis. 149-75. Hershey, PA: Information Science Publishing.

Gaillard, R., Dehaene, S., Adam, C., et al. 2009. Converging intracranial markers of conscious access. PLoS Biology 7(3): e1000061.

Haikonen, Pentti O.A. 2009. Slippery steps towards phenomenally conscious robots. APA Newsletter on Philosophy and Computers 08:2.

Jensen, O., & Colgin, L. L. 2007. Cross-frequency coupling between neuronal oscillations. Trends in Cognitive Sciences 11(7): 267-69.

Massimini, M., Ferrarelli, F., Huber, R., et al. 2005. Breakdown of cortical effective connectivity during sleep. Science 309: 2228-32.

Mitchell, M. 1993. Analogy-making as Perception. Cambridge MA: The MIT Press.

Uchida, N., Kepecs, A., and Mainen, Zachary F. 2006. Seeing at a glance, smelling in a whiff: rapid forms of perceptual decision making. Nature Reviews Neuroscience 7: 485-91.

van Berkum, J. J. A. 2006. Discourse and the Brain. Paper presented at the 5th Forum of European Neurosciences: The Federation of European Neuroscience Societies (FENS), Vienna, Austria.

Willis, J., & Todorov, A. 2006. First impressions: making up your mind after a 100-Ms exposure to a face. Psychological Science 17: 592-99.

Conscious Perception Missing. A Reply to Franklin, Baars, and Ramamurthy

Pentti O. A. Haikonen
University of Illinois at Springfield

Franklin, Baars, and Ramamurthy kindly clarify their position in response to my critique (Haikonen 2009). I fully recognize the important and pioneering work that Franklin, Baars, and Ramamurthy have done in the field of artificial cognition and my critique should not be construed to diminish the value of that work in any way.

However, in the good tradition of philosophical debate I would like to point out the following. There seems to be nothing in the writings of FBR (or anybody else's) that would explain how the running of any computer program could evoke qualia and subjective feelings in the executing machine. On the other hand, it is obvious that computer programs can simulate various feelings including pain and pleasure via their functional consequences. The presence of such consequences does not, however, prove that the computer would actually feel something or be conscious in the hard sense (h-consciousness, see Boltuc 2009). It may well be that the phenomenal aspects of consciousness are beyond the capacity of computer programs and may be present only in some hard-wired perceptive and reactive systems.

FBR have done excellent work in the development of the LIDA agent that they call functionally conscious. Based on that they wish to define functional consciousness as the process that implements the Global Workspace Theory and the LIDA Cognitive Cycle (Franklin, Baars, and Ramamurthy 2009) a notion that I criticized in my previous response. The concept "functional consciousness" is doubtful, but even so, it should not be hijacked to apply to one specific cognitive model only. In their current response FBR wish to go even further. They state: "There is a sizable and growing body of evidence from cognitive science and neuroscience that human minds (their control structures) implement the essential elements of Global Workspace Theory." This is not a modest claim at all.

Personally I would be quite happy if it could be shown that my cognitive model (Haikonen 2003, 2007) had captured some elements of human cognition (I trust it has), but I would not dare to claim that the brain implements my model, even in the unlikely case where my model would turn out to be a perfect model of the brain. The human brain and mind are a little bit older constructions than the Global Workspace Theory and have evolved without any knowledge of the same. It seems that here FBR have switched the role of a natural object and its man-made model. No natural object or system is based on man-made models or blueprints. To claim the opposite is to nominate oneself as the Creator. However, we are free to find Nature's principles and implement those in our own designs.

References

Franklin, Stan, Baars, Bernard J., Ramamurthy, Uma. 2009. A phenomenally conscious robot. APA Newsletter on Philosophy and Computers 08:2.

Boltuc, Peter. 2009. The philosophical issue in machine consciousness. International Journal of Machine Consciousness 1: 155-76.

Haikonen, Pentti O.A. 2009. Slippery steps towards phenomenally conscious robots. APA Newsletter on Philosophy and Computers 08:2.

Haikonen, Pentti O.A. 2003. The Cognitive Approach to Conscious Machines. UK: Imprint Academic.

Haikonen, Pentti O.A. 2007. Robot Brains: Circuits and Systems for Conscious Machines. UK: Wiley & Sons.

ONTOLOGICAL STATUS OF WEB-BASED OBJECTS

A Semantics for Virtual Environments and the Ontological Status of Virtual Objects

David Leech Anderson
Illinois State University

Abstract

Virtual environments engage millions of people and billions of dollars each year. What is the ontological status of the virtual objects that populate those environments? An adequate answer to that question requires a developed semantics for virtual environments. The truth-conditions must be identified for "tree"-sentences when uttered by speakers immersed in a virtual environment (VE). It will be argued that statements about virtual objects have truth-conditions roughly comparable to the verificationist conditions popular amongst some contemporary antirealists. This does not mean that the virtual objects lack ontological standing. There is an important sense in which virtual objects are no less real for being mind-dependent.

Introduction

What is the ontological status of the virtual objects that populate the burgeoning virtual worlds that reside on the Internet? Second Life is a virtual world comprised not only of objects like tables, chairs, trees, and fountains, but large landmasses that have cities with expansive real estate developments. The people who frequent this virtual environment (by animating virtual bodies known as "avatars") not only prize these virtual objects in some Platonic way, they place a commercial value on them by paying cold, hard cash. Anshe Chung (a.k.a. Ailin

Graef) became a cause célèbre and made it on the cover of Business Week magazine when the value of her combined virtual holdings in Second Life exceeded $1,000,000.[1] Not one million in virtual "Linden" dollars (the currency within Second Life), but one million dollars U.S.

World of Warcraft, with over 11 million paying monthly subscribers, is another massively multiplayer online role-playing game (MMORPG) that is played within a virtual environment. The virtual objects that populate games like this are bought and sold earning their creators hundreds of millions of dollars each year.[2] One might begin a discussion about the ontological status of such virtual objects by invoking a famous claim advanced by Ian Hacking, a party to the realism-instrumentalism debate on the ontological status of theoretical entities (like electrons). Hacking reports on a conversation he had with a working physicist. He recounts:

> Now how does one alter the charge on the niobium ball? "Well, at that stage," said my friend, "we spray it with positrons to increase the charge or with electrons to decrease the charge." From that day forth I've been a scientific realist. So far as I'm concerned, if you can spray them then they are real.[3]

In the same spirit, one might be tempted to say: "If you can buy them and sell them for hundreds of millions of dollars then they are real." Even so, merely acknowledging the reality of virtual objects gets us no distance to understanding in what that reality consists. The first step must be to determine what counts as a "virtual object" and then we can ask where (if at all) in our ontological hierarchy they are properly placed.

Virtual environments and virtual objects

From the outset, the meaning of "virtual object" (VO) will be restricted so as to exclude some digital entities that are so described in discussions about Internet commerce. The day of Michael Jackson's memorial service, Facebook gave away 800,000 copies of what was described as a "special commemorative virtual gift"—a graphical representation of Michael's white sequined glove. In this discussion, graphical images like the white sequined glove will not be treated as a virtual object, because a Facebook page does not qualify as a genuine "virtual environment." It lacks nothing in its virtuality, but for the purposes of this discussion it isn't rich enough to constitute an environment, which will be defined as follows:

> E = an environment is a dynamic space-time region (with a minimum of two but typically three spatial dimensions) populated by objects that bear spatial and temporal relations to one another.[4]

The environments which we typically inhabit are physical in nature.

> PE = a physical environment is a dynamic space-time region that consists of objects that bear spatial and temporal relations to one another and objects whose identity conditions include intrinsic, non-relational properties that exist independent of the present cognitive states (thinking, believing, experiencing, etc.) of any cognitive agent.

An MMORPG also constitutes an environment. It provides objects to serve as the target for agents' intentional states and a stage upon which human actions (buying, killing, lying, sharing, etc.) become intelligible and even morally evaluable. We will describe these environments as "virtual" and I propose that we define them as follows:

> VE = a virtual environment is a dynamic space-time region that consists of objects that bear spatial and temporal relations to one another and whose identity conditions supervene on the actual (or possible) sensory and cognitive states of the agents who inhabit that environment.
>
> Virtual environments are populated with virtual objects.
>
> VO = a virtual object is an empirically detectable, intersubjectively stable, publically accessible entity that can be identified and re-identified over a sufficiently long run and whose identity conditions supervene on the actual (or possible) sensory and cognitive states of the agents inhabiting the associated VE.

The reader will notice that the definitions of VE and VO make their very existence dependent upon the cognitive agents who engage with them. If a plague wiped out the entire human race, and there no longer existed agents for whom the virtual tree could be an object of experience, then the tree would cease to exist.

An opponent will object that this consequence is unreasonable. Virtual environments (VE's) need not be these ephemeral things that blink out of existence when the cognitive agents who previously inhabited them cease to exist. VO's could, instead, be defined so as to be constituted by the physical systems that are causally responsible for generating users' experiences of them. In that case, even if all humans died in a plague, so long as the machinery kept running, the VO's would continue to exist.

Those sympathetic to this latter interpretation of VO's will object that my previous definition attempts to settle, by stipulation, what should be a central controversy of this paper. This is a fair objection. I concede that my definitions of VE and VO do require a convincing argument—an argument that will be offered below. Laying these admittedly contentious definitions on the table at the outset will, however, streamline the discussion and aid in the explication of the view. In the end the reader must judge whether the definitions are well motivated.

Virtual Environments (VE's) in the history of philosophy

By making virtual environments dependent upon the cognitive activity of the agents inhabiting them, I am quite deliberately invoking the idealist / antirealist philosophical traditions and the various alternatives they offered to a traditional realist account of the nature of the external world. Consider how the language of VE's and PE's might be used to describe debates over realism from Descartes to the present. We might recast Descartes' methodological skepticism by considering the possibility that while I am having the experiences of a tree in the quad (VO), there may not exist a physical tree (PO) that answers to it. We could then imagine Berkeley arguing that the very concept of a material tree (PO) is incoherent and all that normal people mean when speaking of a "tree" is the tree-as-experienced (the VO). Michael Dummett's "language acquisition argument" will translate into the claim that there is no coherent account of how humans could learn to understand a language with realist truth-conditions (which requires asserting the existence of PO's) and so the correct semantics for natural languages must be verificationist (which requires asserting the existence of VO's only).[5] Finally, Hilary Putnam's "Brains in vat" argument can be seen as an attempt to show that even if one begins with externalist assumptions, one will utterly fail in one's attempt to raise the Cartesian specter of radical skepticism by considering the case where I am a disembodied brain floating in a vat of nutrients stimulated by a computer (experiencing VO's but not PO's). Ironically, knowing that I am not a brain in a vat does not confirm the truth of realism, according to Putnam.[6] On the contrary, if I can't raise the specter of radical skepticism—because there is no genuine possibility that I could be in a VE that is not being caused by a corresponding

PE—then the very distinction between VE's and PE's collapses, as most antirealists insist, and we are left with the incoherence of metaphysical realism.[7]

Admittedly, I am playing fast-and-loose with these august traditions as I reformulate them in terms of VE's and VO's. But I do hope that in spite of any quibbles one might have about my reading of history, the general point can still be made. The territory we are exploring is not unique to the twenty-first-century world of MMORPG's. And the decisions we make about the proper analysis of VE's might conceivably speak to or even commit one to certain positions in much broader areas of philosophy.

One more comment before we progress. The reader should be cautioned not to conclude that embracing a verificationist semantics for virtual objects in any way counts against the truth of metaphysical realism regarding the external world. Quite the contrary is actually the case, or so I would argue if space permitted.

Ontology with semantics

Our ultimate goal is to determine the ontological status of virtual objects. One cannot determine the ontological status of a particular virtual tree, however, unless one first determines what the virtual tree is. But that is ultimately related to the semantic question, What are the truth-conditions of the sentence, "There is tree" when spoken about the virtual tree in the virtual environment? The point is not that you can ignore metaphysics and simply do semantics. Certainly not. The point is that reality is too metaphysically rich; there are simply too many realities that are prima facie candidates for being the referent of the virtual "tree" in question. Nothing will be accomplished if one gives the most elegant ontological account of a phenomenon that one takes to be a virtual tree, if everyone in your audience insists that the phenomenon that you analyzed is simply not a "virtual tree" given what everyone else means by those words. The war is only won if you win both the semantic battle and the ontological battle.There is no avoiding the semantic question. We want to know the ontological nature of virtual trees; but to answer that question, we must also learn what the truth-conditions are for "tree"-sentences uttered within virtual environments. It is to both questions that we now turn.

A slippery slope argument

When native English speakers are immersed in an MMORPG and utter a "tree"-statement, the language they are speaking (whatever language that turns out to be) we will call "V-English." I have proposed that the truth-value of such "tree"-sentences is not sensitive to the state of any physical object and is instead determined solely by how things stand with respect to the cognitive states of the inhabitants of the game. Thus, "tree"-sentences will be true on this account if all or most inhabitants of the game are having cognitive experiences constitutive of an "empirically detectable, intersubjectively stable, publically accessible" tree.

An alternative to this view is one motivated by the currently very popular position of semantic externalism. On one version of this view, a word refers to whatever it is that "causally regulates" the use of that term. The referent of the term is whatever it is that lies at the end of the causal chain that ultimately causes the speaker to utter sentences like, "Look, there is a tree." For a speaker immersed in a particular MMORPG, there are a number of candidates that might fit that description. Let's consider a few. Each of the numbered conditions that follow is a candidate for being that physical state-of-affairs that causally regulates the use of the term, "tree," when uttered by a person immersed in a VE. When identified, that condition will be the referent for the term, "tree," in V-English. The first candidate is:

> 1. States of the server hardware: The throwing of the electrical (on-off) switches on the VE-server that implements the "tree"-subroutine in the server software.

This has been a popular choice for an externalist referent of "tree"-statements as uttered by "brains in a vat" according to Putnam's thought experiment,[8] and at first blush seems equally promising here. The software alone is hard to target because it is an abstract entity, not a physical object. The switches alone, outside of the context provided by the "tree"-software, don't capture the continuity of the tree through time. This articulation attempts to capture the best of both worlds, embracing both the server's hardware and software. But this option is susceptible to a counterexample.

Assume that a bug is detected in the VE-server software of a famous MMORPG right in the middle of a well-publicized national contest being waged live and online between just two contestants. The two will soon be shooting (virtual) arrows into a tree and the bug can be expected to produce a malfunction. In order to prevent this eventuality, the programmer has a plan. The programmer knows exactly what signals (plus TCP/IP – Internet protocols) would be sent out over the Internet to the two contestants' PC's if the server were functioning properly. Imagine, as well, that the programmer has the ability (it doesn't matter how) to interrupt the stream of defective instructions whenever they are sent by the software bug, and to send instead packets of instructions over the Internet that produce the proper "tree"-effects on the users' computers. (Consider the programmers actions here on analogy with the actions of God in correlating the actions of minds and bodies according to "Occasionalist" theories of mind-body interaction.)

In this case—where the programmer is ensuring that inhabitants of the VE will continue to experience an "empirically detectable, intersubjectively stable, publically accessible" tree—what do we say about the truth-value of the sentence, "The tree was struck by an arrow" when uttered within the MMORPG? It seems only reasonable to say that it is true. The quick-thinking, spontaneous actions of the programmer preserved the existence of the tree within the VE. However, if the V-English word, "tree," refers to condition 1. identified above, then the sentence must be false. Because of the software bug, the "tree"-routine is no longer being run. But "false" is the wrong result. As a matter of fact, it turns out neither hardware nor software is either a necessary or a sufficient condition for the existence of the virtual tree. Happily, the case itself suggests a second (PO) candidate for the referent of the word "tree":

> 2. Signals propogated over the Internet: The carrying of the instructions plus TCP/IP Internet protocols that propogate the "tree-generating" instructions sent to the personal computers of all agents inhabiting the VE.

The reader herself, however, can probably generate a counter-example to this proposal. Instructions sent by TCP/IP is again just a half-way house. It is only a means for delivering instructions to the PC's of each participating agent. There are any number of methods that might accomplish this, including tens of thousands of employees scattered around the world using all manner of quirky occasionalist methods for getting signals to all relevant computers causing the proper pixels to light up and forming the 3D image of a tree. So long as the result is a stable, intersubjectively consistent, genuinely public, virtual tree—"tree"-statements uttered in V-English will still come out true, even when condition 2. is lacking.

But now we are on a slippery slope. "Pixels lighting up on computer monitors around the world" is no more the proper "end" of the causal chain than any of the previous ones. Pick

any condition on the list. It is possible for that condition not to hold but so long as the next condition down on the list does hold, then the existence of the virtual tree will be preserved.

> 3. The hardware in every user's PC: The aggregate throwing of all appropriate electrical (on-off) switches in the machine hardware that implements the "tree"-subroutine on all the personal computers of all the agents inhabiting the VE.
>
> 4. The monitor illumination in every user's PC: The synchronized illumination of pixels on all the computer monitors of all agents so as to create an intersubjectively consistent 3D public tree.
>
> 5. Retinal stimulation in every user's eyeball: The proper stimulation of the retinas in all agents' eyes so as to create an intersubjectively consistent 3D public tree.
>
> 6. Visual cortex stimulation in every user's brain: The excitation of the proper areas of the visual cortex (V1–V5) in all agents' brains so as to create an intersubjectively consistent 3D public tree.

Each of the conditions is neither necessary nor sufficient for the existence of the virtual tree. It doesn't matter how you accomplish the task of bringing about the conditions described by VE and VO. It doesn't matter what physical system (or mental or spiritual system for that matter) is used along the way, all that matters is that you produce the final effect that literally constitutes the existence of the virtual tree. And that final effect inevitably is[9]:

> 7. The intersubjectively coordinated, conscious experience of all users: The proper production of an empirically detectable, intersubjectively stable, publically accessible tree that can be identified and re-identified over a sufficiently long run and whose identity conditions supervene on the sensory and cognitive states of the agents inhabiting the associated VE.

When typical inhabitants of a MMORPG confront a tree, and say, "There is a tree," what they are talking about is best captured by 7.

The ontological status of virtual objects

I have just argued that the essential nature of virtual trees is best described not in terms of the mind-independent states of physical systems, but in terms that make essential reference to the cognitive states of human agents. One might reasonably use either the language of a verificationist semantics as embraced by some antirealists, or the language of conscious, first-person phenomenal states familiar from recent defenders of phenomenal consciousness. I have purposely used both in this discussion, not wanting to privilege either one. Some will likely find one option more congenial than the other.

I recognize that I have done nothing to answer the myriad objections that can quite reasonably be raised against this controversial position. But that work must be left to another time. Our final task now is to come full circle on the Hacking quote that opened this paper. If (by chance) I am right about the essential nature of virtual objects, are they real or not? Do we add them to our ontology, or leave them out?

Lynn Rudder Baker has, in the pages of this publication (Spring 2008[10]), addressed a question that is, at least in part, relevant our question: Are artifacts less real than natural objects because they are mind-dependent? In that article, she defends a position with which I am completely sympathetic. She insists that human artifacts are in no way metaphysically deficient in virtue of their having been shaped and fashioned by the human mind. Tables are no less metaphysically "real" than are tubers. She says,

> There is a venerable—but, I think, theoretically misguided—distinction in philosophy between what is mind-independent and what is mind-dependent. Anything that depends on our conventions, practices, or language is mind-dependent (and consequently downgraded by those who rest metaphysics on a mind-independence/mind-dependence distinction)...
>
> A second reason that the mind-independent/mind-dependent distinction is unhelpful is that advances in technology have blurred the difference between natural objects and artifacts. For example, so-called "digital organisms" are computer programs that (like biological organisms) can mutate, reproduce, and compete with one another. …Are these objects… artifacts or natural objects? Does it matter? (p.4)

I wholeheartedly affirm Rudder Baker's sentiments here. A chair is not dependent upon the present cognitive activity of any agent. Every mind that exists in the universe could cease to exist and the chair would continue to exist. The chair is causally dependent upon the past activity of some cognitive agent, but so long as that past activity produced something with its own intrinsic properties, which is dependent upon no present mental activity, then it is mind-independent in the metaphysically relevant sense.

Those who Rudder Baker criticizes make the mistake of conflating two fundamentally different meanings of the term "mind-dependence." The kind of mind-dependence attributable to artifacts which are dependent only in their causal origins has profoundly different ontological implications than the kind of mind-dependence attributable to virtual objects which will literally cease to exist if minds stop thinking about them. Treating the former as if it deserves the same ontological classification as the latter ignores this important distinction. That is why, according to my definitions, artifacts qualify as physical objects (PO) not virtual objects (VO).

Having said that, I would argue that Rudder Baker makes too sweeping a claim in her concluding two sentences of the paper, where she says:

> No one who takes artifacts of any sort seriously, ontologically speaking, should suppose that metaphysics can be based on a distinction between mind-independence and mind-dependence. In any case, technology will continue to shrink the distinction, and with it, the distinction between artifacts and natural objects. (Ibid.)

Ironically, Rudder Baker seems to be committing the same conflation error as did her opponents, but in the reverse direction. She seems to be denying that there could be any ontologically significant distinction between mind-dependence and mind-independence. The table in my kitchen, and the table in my Second Life kitchen are both artifacts. But one is mind-dependent in a more robust sense that is ontologically significant. If she is denying this, then I think she makes a mistake similar to her opponents. In the end, I am not at all confident that she is denying this distinction. She may simply be ignoring it. But if she is denying it, then we deserve more of an argument than she has given thus far.

Conclusion

So what about virtual trees? Are they real? The term "real" is ambiguous. No, they are not real in the sense that is opposed to being ideal, or verificationist. They are metaphysically dependent upon the cognitive state of human beings, and that makes them "ideal" in contrast to physical trees that are "real."

But more important than being "real," in this sense, is being real in the sense that they "exist" in some substantive sense of that word. Yes, they do exist. They are real enough to make it the case that statements like "I left your shield next to the tree in the quad" and "You lied to me about the power of this sword and I will never be your friend again" are true. These statements are not about a fictional realm like that of a novel. One can commit real (not fictional) betrayal with a virtual sword, and a sword real enough to betray is real indeed.

I also believe it is reasonable to say that virtual trees exist and should be listed among our ontological commitments in something like the way that conscious states should be listed in one's ontology.[11] The conscious states of John's believing that p and seeing an orange sunset are real. So too, the virtual tree that supervenes on the conscious states of many people. Obviously, those who deny the existence of consciousness will not be persuaded by this line of reasoning, but for that there is also the verificationist route to metaphysical legitimacy.

This discussion has only begun to explore the ontological status of virtual objects. It is a discussion that I hope continues.

Endnotes

1. May 1, 2006, Business Week cover title "Virtual World, Real Money."
2. To learn more about the very real business of buying, selling, and owning virtual objects, see the online newsletter, Virtual Goods News, http://www.virtualgoodsnews.com.
3. Ian Hacking. Representing and Intervening (Cambridge: Cambridge University Press, 1983), 23.
4. I don't mean to engage any philosophical disputes surrounding the nature of "space-time," if only because I have no competence in that domain. I am hoping for a conception as neutral as is possible, believing that nothing of significance hangs on it.
5. Michael Dummett. Truth and Other Enigmas (Cambridge, MA: Harvard University Press, 1978).
6. Hilary Putnam. Reason, Truth & History (Cambridge: Cambridge University Press, 1983), 1-21.
7. For more on the arguments of Putnam and Dummett and their relationship to the realism-antirealism debate, see David L. Anderson, "What is Realist about Putnam's Internal Realism," Philosophical Topics 20 (1992): 49-84.
8. Ibid.
9. Those who embrace the Identity thesis for the nature of mental states might want to argue that condition 6. more accurately describes the very same reality as condition 7. Since I reject the Identity thesis, I don't hold this view but in the present context I wouldn't feel compelled to argue with such a person. Condition 6. is close enough to victory for me.
10. Lynn Rudder Baker. "The Shrinking Difference between Artifacts and Natural Objects." American Philosophical Association Newsletter on Philosophy and Computers 07 (Spring 2008): 2-5.
11. For more on the reality of conscious, see David Leech Anderson, "Consciousness & Realism" Journal of Consciousness Studies 14 (2007): 1-17.

Realism and Antirealism in Informatics Ontologies

Robert Arp
National Center for Biomedical Ontology, University at Buffalo

Abstract

The realism-antirealism debate in the philosophy of science has made its way into informatics and computer science circles in debates concerning the status of the entities represented in what informaticians call ontologies. For realists, the terms of these ontologies refer to real entities out there in the world; for antirealists, they refer to concepts in the minds of experts. In this paper, after an explanation of domain and formal ontologies, I offer some criticism of the antirealist approach and argue that, in spite of the antirealist sentiments that still predominate in informatics circles, informaticians can nonetheless feel comfortable in constructing domain and formal ontologies from a realist perspective.

Key Words: informatics, ontology, domain ontology, formal ontology, realism, antirealism

The Sea of Information

Informatics is the science associated with the collection, categorization, management, storage, processing, retrieval, and dissemination of data and information—principally, through the use of computers as well as computational and mathematical models—with the overall goal of improving retrieval and dissemination of data and information. Increasingly, many more traditional disciplines have their own informatics, reflecting the fact that they are confronted by the need to deal with large bodies of data and information—consider, for example, the field of Geographic Information Systems (http://www.gis.com/) or of biomedical informatics (Shortliffe & Cimino 2006).

The body of information deriving from such disciplines that is now being made freely available through computers on the Web constitutes a veritable sea of extraordinary depth and breadth. How can we collect, categorize, manage, store, process, retrieve, disseminate, mine, and query all of this data and information appropriately and efficiently by computational means?

The problem is to chart this ever-growing sea of information in such a way that its various parts can be efficiently accessed, used, navigated, and reasoned about by human individuals. How can we ensure that the terminology, definitions, relations that are used when storing information and data (a) accurately reflect the developing state of knowledge in a particular domain or discipline, (b) are internally coherent, (c) are clearly defined, and (d) are interoperable from one database to the next?

Here, it is especially (d) that poses problems. Researchers in different disciplines speak different languages, use different terminologies, and format the results of their research in different ways. The situation is not unlike that of the Biblical Tower of Babel, where there is an uncontrolled and unsurveyable multiplicity of different languages and little in the way of cross-linguistic understanding. Because bodies of data are insulated from each other in this way, interoperability, shareability, and reusability of data and information is greatly limited. The result is a silo effect: data and information are isolated in multiple, incompatible silos. And it is to address the silo problem, philosophers, computer scientists, and informaticians in various domains have worked to create what are known as domain ontologies in their respective fields of study.

What is a Domain Ontology?

A domain is a larger or smaller area, sphere, aspect, or delineated portion of reality, such as geography, ecology, law, or toxiconutrigenomics—which humans seek to know about, understand, and explain (also possibly, predict, manipulate, and control) as fully as is possible through the development of a corresponding science or discipline. An ontology is a little more complicated to define.

Traditionally, of course, the word "ontology" refers to a branch of Western philosophy that has its origins in Greek metaphysics and is concerned with the study of being. From this philosophical perspective, ontology seeks to provide a definitive and exhaustive classification of entities in all spheres or domains of being. Thus, the Random House College Dictionary (2007) defines ontology as "the branch of metaphysics that studies the nature of existence."

Since the emergence of the information age, the term "ontology" has also come to be used by computer scientists to refer to structured representations of the entities and relations existing within particular domains of reality (Gruber 1993). Ontology, in the philosophical sense, has all of reality as its subject matter. A domain ontology, by contrast, is a controlled, structured vocabulary for a specific discipline—providing a backbone taxonomy or hierarchy of types and subtypes and a set of logically defined relations between its terms. It thereby serves as a framework for the annotation of data within a particular domain. Its purpose is to make the data in the corresponding discipline more easily searchable by human beings and more efficiently and reliably processable by computers (Smith 2003). The ontology is designed also to ensure that the different bodies of data collected by different researchers in the same domain should all be represented in the same way, and in this way it serves to counteract the formation of silos. Fully developed domain ontologies are utilized by researchers especially in the biomedical field. Examples include:

Common Anatomy Reference Ontology (CARO)

Foundational Model of Anatomy Ontology (FMAO)

Environment Ontology (EnvO)

Infectious Disease Ontology (IDO)

Ontology for Biomedical Investigations (OBI)

Phenotypic Quality Ontology (PATO)

Cell Ontology (CL)

Sequence Ontology (SO)

Protein Ontology (PRO)

all of which are part of the Open Biomedical Ontologies Foundry (http://obo foundry.org/; also Smith, Ashburner, Rosse, Bard, Bug, Ceusters, et al. 2007).

Formal Ontology

There are many factors that contribute to silo effects, including poor conceptualizations, faulty linguistics, and the need to deal with layers of information deriving from multiple different and independent sources. A further problem, however, is that the very success of the strategy of creating domain ontologies in order to counteract these effects has led to the creation of multiple special-purpose domain ontologies, which have served to re-create the silo effect at a new level. To remedy this problem, a third kind of ontology—a formal ontology—has emerged, with the goal of constraining domain ontologies in such a way that they satisfy certain basic principles, including logical principles, in a way which maximizes interoperability. In fact, this is where philosophers have made important contributions to informatics by pointing out the pitfalls of poor reasoning that have hampered information accessibility and dissemination thus far. The ultimate, and still visionary, goal of formal ontology is the calibration of all domain ontologies into one single, organized, interconnected, and interoperable computer repository, accessible in real time to anyone anywhere in the world.

Many people use the word "formal" as interchangeable with "upper-level," "top-level," or "higher-level" and, indeed, this is appropriate since formal ontologies assist in making communication between and among "lower-level" domain ontologies possible by providing a common formal framework or ontological backbone. Stated simply: interoperability of domain ontologies is more likely to occur, to the degree that researchers are using the same upper-level ontological categories and relations. Existing upper-level or formal ontologies include:

- Standard Upper Merged Ontology (SUMO) of the Institute of Electrical and Electronics Engineers (http://suo.ieee.org/),
- Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (http://www.loa-cnr.it/ DOLCE.html), and
- Basic Formal Ontology (http://www.ifomis.org/bfo);

and work is currently underway to create a merger of these three in a new consensus formal ontology for future use.

Realism and Antirealism in Ontologies

There is an age-old debate from philosophical ontology that has made its way into the field of domain and formal ontology known as the realism-antirealism debate. This debate can be traced back to Socrates and Protagoras. Socrates believed that there was a real world of universals and particulars "out there" beyond the mind's influence; Protagoras essentially denied this in favor of various forms of what we can now recognize as subjectivism, conceptualism, constructivism, and/or idealism. There are many different types of realism and antirealism discussed in philosophy today disciplines (Brock & Mares 2007; Alston 2002; Kulp 1997; Papineau 1996). Here, I am dealing with the kinds of realism and antirealism discoverable in informatics circles, specifically concerning scientific domain ontologies.

According to one common version of realism defended in these circles:

- there is a reality independent of the mind's awareness of it;
- reality is what it is (at least in domains outside psychology and engineering), independently of the mind's influence;
- the representational units of domain and formal ontologies represent real entities and the relationships between them.

The biomedical ontologies in the OBO Foundry referred to above are being developed on the basis of a realist perspective of this sort, as expressed, for example, in Rosse and Mejino (2003, 2007). From the realist perspective, a domain ontology is a representational artifact whose representational units are intended to designate the universals and relations in a given domain. Such an ontology corresponds, in effect, to that part of the content of a scientific theory that is captured by its constituent general terms (Smith 2004; Smith, Kusnierczyk, Schober, & Ceusters 2006).

A corresponding standard version of antirealism denies or transforms each one of the theses above:

- there is no reality independent of the mind's awareness;
- reality is wholly a product of the mind's influence;

- the representational units of domain and formal ontologies represent our concepts, beliefs, ideas, etc., of so-called "real world entities."

Researchers such as Corcho, Fernandez-Lopez, and Gomez-Perez (2006), Beynon-Davies (2003), Frank (2007), and McCray (1993, 2006; also Gruber 1993) subscribe to one or other version of this view. Corcho, Fernandez-Lopez, and Gomez-Perez (2006) define an ontology as a "formal, explicit specification of a shared conceptualization" where "conceptualization" refers to "an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon" (p. 4). Further: "[An] ontology is a theory of what entities could exist in the mind of a knowledgeable agent" (VanHeijst, Schreiber, & Wielinga 1996). Consider, also, Corcho and Gomez-Perez (2000): "Concepts, also known as classes, are used in a broad sense. They can be abstract or concrete, elementary or composite, real or fictious. In short, a concept can be anything about which something is said, and, therefore, could also be the description of a task, function, action, strategy, reasoning process, etc." (p. 81).

The general and prevalent sentiment of antirealism among informaticians is captured in the following definition of ontology from the website, owlseek.com:

> We can never know reality in its purest form; we can only interpret it through our senses and experiences. Therefore, everyone has their own perspective of reality. An ontology is a formal specification of a perspective. If two people agree to use the same ontology when communicating, then there should be no ambiguity in the communication. (Smith 2004, 32)

So, in informatics circles, there are two general camps of realists and antirealists—existing in various stripes—who are increasingly engaging in debate. Examples of the sorts of claims one hears from the antirealist side, which still predominates in such debates, are:

- "…we all perceive things differently, so the best we can do is come to the table and agree on the meanings of terms in our ontologies."
- "…but, how can you really know reality? I'll use whatever ontology gets the job done."
- "…your formal ontology is a gold standard, I'll grant that; but a gold standard is just a really big consensus agreement."

Problems with Antirealism

There are problems with antirealism, however, as captured in the claims such as those quoted above. These problems are all the more poignant when we consider domain ontologies that are constructed for scientific research, as opposed to domain ontologies created to support manufacturing or administration purposes, or for domains of imagined objects such as fiction or mythology. These problems, perhaps best summarized by David Stove here: http://www.maths. unsw.edu.au/~jim/worst.html, are familiar to philosophers, but seem not to have penetrated informatics circles.

First, there must be some things outside of our minds; otherwise people would walk off cliffs more often, not stub their toes when kicking rocks, and never go to the doctor because "the surgery will never work…it's all in my mind, anyway."

Second, we would never be able to predict, with any degree of reliably, that a storm is coming, that antibiotics will assist in killing the bacteria, or that a bridge will support the weight of your car—but, obviously, we do make such reliable predictions. In fact, it would be quite miraculous if our perceptions, concepts, and beliefs did not match up with some objective reality, given the general reliability of our perceptions, concepts, and beliefs (Sankey 2006).

Third, and more to the point concerning scientific domain ontologies, it seems absurd to think that there is no order or structure to a reality that is independent of us, our language use, and our conceptualizations. Domain ontologies in the scientific realm track reality, and take the forms that they do at least in part as a result of the fact that reality is structured the way that it is. Before homo sapiens evolved there was already a biosphere categorizable at various levels of granularity into cells, organisms, populations, and beyond. Does anyone really think that this categorizability is dependent upon some mind or minds who evolved only some 65 million years later?

When it comes to scientific domain ontologies, science ultimately concerns itself, not with particular instances of things but with their essential natures, with universals or types. Once data regarding individual particulars are gathered, there is a natural process of sorting the data into categories—not of the particulars themselves—but of the universal features that the particulars share. For example, the genus Felis is an actual universal that scientists track, and believe to exist, not by virtue of your cat Fritz, the specific leopard that your zoologist brother has been studying, or the ocelot they have in Chicago's Lincoln Park Zoo, but rather by virtue of the universal feature or characteristic shared by all of these individual, particular instances of felines.

Further, in science and in our everyday lives we want to know not about the concept, belief, or perception concerning something, but about the actual thing itself, irrespective of anyone's subjective or intersubjective conception, belief, or idea of it (Rescher 2005). We seek the actual facts to support our hypotheses, corroborate our theories, or legitimize the data that is imported into our computational systems.

Finally, antirealism and its various subvarieties of subjectivism, conceptualism, constructivism, idealism, and so forth is an entertaining idea to think about, but the actual work done by scientists in the investigation of entities like tumors, Amyotrophic Lateral Sclerosis, stellar cell, and mouse kidney seems for the most part to presuppose a realist ontology and methodology (Leplin 1997).

Despite the antirealist sentiments found in informatics circles, informaticians can feel comfortable to construct domain and formal ontologies from a realist perspective—especially if the ontologies on which they are working emerge from the sciences. Of course, if the domain ontology is specifically about fictional characters or mental makeshifts like myths or monsters, then the representational units of the domain ontology will refer to concepts of minds (consider, for example, the ontology behind the role-playing game Dungeons & Dragons: http://www.wiz ards.com/). However, the entities to which scientific ontologies refer can be real entities that exist out there in the real world, and not merely concepts and the like.

*This work is funded by the United States National Institutes of Health (NIH) through the NIH Roadmap for Medical Research, Grant 1 U54 HG004028. Information on the National Centers for Biomedical Computing can be found at http://nihroadmap.nih.gov/bioinformatics.

References

Alston, W. (Ed.). 2002. Realism and Antirealism. Ithaca, NY: Cornell University Press.

Beynon-Davies, P. 2003. Information Systems: An Introduction to Informatics in Organizations. New York: Macmillan.

Brock, S., and Mares, E. 2007. Realism and Antirealism. Montreal: McGill-Queen's University Press.

Ceusters, W., and Smith, B. 2006. A realism-based approach to the evolution of biomedical ontologies. Proceedings of AMIA Symposium 2006 1: 121-25.

Corcho, O., and Gomez-Perez, A. 2000. A roadmap to ontology specification languages. In Knowledge Engineering and Knowledge Management: Methods, Models and Tools, edited by R. Dieng and O. Corby. 80-96. New York: Springer.

Corcho, O., Fernandez-Lopez, M., and Gomez-Perez, A. 2006. Ontological engineering: Principles, methods, tools and languages. In Ontologies for Software Engineering and Software Technology, edited by C. Calero, F. Ruiz, and M. Piattini. 1-48. New York: Springer.

Frank, A. 1996. Ontology: A consumer's point of view. In Spatial and Temporal Reasoning, edited by O. Stock. 135-54. Dordrecht: Kluwer.

Frank, A. 2007. Towards a mathematical theory for snapshot and temporal formal ontologies. Lecture Notes in Geoinformation and Cartography 1863-2246: 317-34.

Gruber, T. 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5: 199-220.

Kulp, C. (Ed.). 1997. Realism/Antirealism and Epistemology. New York: Rowman & Littlefield.

Leplin, J. 1997. A Novel Defense of Scientific Realism. Oxford: Oxford University Press.

McCray, A. 1993. Representing biomedical knowledge in the UMLS semantic network. In High Performance Medical Libraries: Advances in Information Management for the Virtual Era, edited by N. Broering. 44-55. New York: Meckler Books.

McCray, A. 2006. Conceptualizing the world: lessons from history. Journal of Biomedical Informatics 39: 267-73.

Papineau, D. (Ed.). 1996. Philosophy of Science. Oxford: Oxford University Press.

Rescher, N. 1997. Objectivity: The Obligations of Impersonal Reason. Notre Dame, IN: University of Notre Dame Press.

Rescher, N. 2005. Commonsense: A New Look at the Old Philosophical Tradition. Marquette: Marquette University Press.

Rosse, C., & Mejino, J. 2003. A reference ontology for biomedical informatics: The foundational model of anatomy. Journal of Biomedical Informatics 478-500.

Rosse, C., and Mejino, J. 2007. The foundational model of anatomy ontology. In Anatomy Ontologies for Bioinformatics: Principles and Practice, edited by A. Burger, D. Davidson, and R. Baldock. 59-118. New York: Springer.

Ruiz, M., Bodenreider, O., Little, E., and Srinivasan, P. 2006. Ontological research and its applications to the biomedical domain. Proceedings of the American Society for Information Science and Technology 42: 37-57.

Sankey, H. 2006. Why is it rational to believe scientific theories are true? In Rationality and Reality: Conversations with Alan Musgrave, edited by C. Cheyne and J. Worrall. 109-32. Dordrecht: Springer.

Shortliffe, E., and Cimino, J. (Eds.). 2006. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. London: Springer.

Smith, B. 2003. Ontology. In Blackwell Guide to the Philosophy of Computing and Information, edited by L. Floridi. 155-66. Malden, MA: Blackwell.

Smith, B. 2004. Beyond concepts: ontology as reality representation. In Proceedings of FOIS 2004. International Conference on Formal Ontology and Information Systems, edited by A. Varzi and L. Vieu. 31-42. Amsterdam: IOS Press.

Smith, B., Kusnierczyk, W., Schober, D., and Ceusters, W. 2006. Towards a reference terminology for ontology research and development in the biomedical domain. Proceedings of KR-MED 1: 1-14.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. 2007. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25: 1251-55.

VanHeijst, G., Schreiber, A., and Wielinga, B. 1996. Using explicit ontologies in knowledge-based systems (KBS) developments. International Journal of Human and Computer Studies 46: 293-310.

DISCUSSION 2: ON FLORIDI

A Response to Barker

Ken Herold
Hamilton College

John Barker ("Too Much Information: Questioning Information Ethics" APA Newsletter, 08/1 (Fall 2008): 16-19) did not pursue his interpretation of informational entities as so-called arbitrary objects along the new ontological path framed by Floridi. Justifying their subsequent rejection as merely ubiquitous, Shannon-like measurements impervious to any natural account of moral action, Barker misses Floridi's plainly-stated response to the objection that he is asserting the moral worth of mere data. Although he usefully associates Floridi's Informational Structural Realism with Information Ethics (IE) and even praises it, he summarily rejects its interpretive implications (e.g., the method of Levels of Abstraction) and leaves an impression that any data object is isolated or is automatically a candidate for information-hood. Floridi's ontological impetus stems from recent developments in philosophies of logic and information, characteristic of the convergence of computer science, in its desire to understand the dynamical properties of state transitions and program behavior, with the quintessential mathematics underlying current exploratory methods for modeling knowledge. Thus, we have logics for information update, Floridi's own logic of being informed, works on logical pluralism and semantic information, in short, a vortex of effort evolving around epistemic logic, doxastic logic, dynamic or active logics, communication logic, and temporal logic. Herein I address the criticism of "too much information" resulting from over-application of a quantitative ontological view by directing the reader to an alternative and new, qualitative perspective.

One fundamental point of departure for understanding the qualitative difference of ontology for IE is Floridi's early reliance on Uexküll and Sebeok's concept of Umwelt in re-crafting the notion of entropy. Floridi and Sanders originally distinguished IE as a non-standard Ethics: as in theories of nature and space, such as Medical Ethics, Bioethics, and Environmental Ethics, not limited by history (human actions) or by time (their consequences). As a culmination of this development of a distinguished domain of action, Computer Ethics itself is seen as moving beyond a biological boundary into the cybernetic, wherein information is acknowledged as the ontological focus of moral worth. In the process, the former imagined, reasonable, and informed self of standard Ethics is transformed, allowing a freedom of moral concern from the biases of an egocentric and anthropocentric agent. The resulting being or informational entity inhabits the infosphere (akin to our human Umwelt or ecological niche), a realm structured by systems whose complexity engenders a new appreciation of entropy suggested by developments in modal theory for computer science. Floridi's infosphere, as a proper notion for such a generalization of focus, may be placed in the philosophical-historical context of Vernadsky's scientific popularizing of the Le Roy-Teilhard noosphere, or Popper's World 3, or Dennett's meme-based infosphere.

Barker narrowly argues against this essential transition of focus of moral worth to a minimally common ontological status among informational entities solely because it is not reasoned upon "a specific account of what constitutes benefit or harm." Has Floridi in fact neglected to address this crucial aspect of IE? Not at all. Given the qualitative ontological shift suggested

above, one finds explicit contributions to the corresponding rationale within the context of mathematical modal logic, in Floridi and Sanders' fundamental requirements for Information Ethics:

* Stability—specifying the state spaces and actions of agents based upon their mathematical properties;
* Modularity—reflecting information system design and supporting incremental reasoning;
* Rigorousness—applying formal methodology and logical consistency in hierarchical reasoning;
* Soundness—respecting the empirical grounding of ordinary judgment and codified values.

Floridi later describes the four fundamental principles (laws) of Information Ethics:

0. entropy ought not to be caused in the infosphere (null law);
1. entropy ought to be prevented in the infosphere;
2. entropy ought to be removed from the infosphere;
3. the flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their properties.

The null law states "do no harm" and its neutrality is the minimal conduct described. (The English passive voice construction indicates patient-orientation and is convertible to an agent-oriented phrasing.) Laws 1 and 2 progressively intervene in providing increasing levels of help, while law 3 embraces the fullness of IE. Floridi in even later writings restates law 3 and relaxes the relative triumph of any one law over another:

3a. the flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their properties.

Initially, Floridi and Sanders conceptualize entropy as a mathematical structure in the objective portion of the infosphere known as cyberspace. It is in this respect that the fundamental requirements for IE are justifiably employed, the moral entropy laws evaluated, and what is neutral (benign), harmful (evil), and helpful (good) defined. This new kind of entropy is ontological rather than syntactic. If we understand thermodynamic entropy and communications theory entropy both as mathematically quantitative as expressed through the regime of symbol combinatorics and ensembles, this informational entropy is philosophically qualitative and articulated through the reiterated development of formal logics. If the bridgework between requirements and laws within the building of IE is founded upon novel applications of computation in philosophy, then it is still unfair to claim it is unfounded or, at best, not founded upon reason, as Barker argues.

If we assume that these moral laws of benefit and harm hinge on understanding Floridi's non-thermodynamic and non-statistical usage of the term entropy, one could re-interpret Barker's objection to be critical of Floridi for having pioneered a new philosophical basis for the concept of entropy with insubstantial reason. This subject is ripe for investigation, I believe, and I will conclude with some research observations.

True, Floridi's notion of entropy is quite distinct from Boltzmann, Shannon-Weaver, von Neumann, Jaynes, Brooks-Wiley, Gatlin, or even Collier. Closer in kind is Chaitin's algorithmic view of entropy, centering on the program required to specify an ensemble of symbols, or an analysis of epistemology as information theory. Another bridging concept towards the qualitative might be Szilard's insight into entropy-memory-intelligence linkages for physical systems. Floridi leap-frogs the treatment of programs as mathematical objects to concentrate instead on universal logical consistency in IE's analysis of an ensemble of informational structures, as it comprises the infosphere. Infosphere is the informational environment in toto, the entire domain of reality, including, but not limited to organized knowledge in its new form as digital macrocosm. The semantic and ontological values of entropy and information are inversely related: an entropy increase degrades meaning and lessens the richness of content, and lowers the amount of information in a diminished infosphere. Perhaps this entropy can be thought of as an order of logical dimension along an axis of knowledge-production. Think of a metaphysical entropy, a Kantian-Quinean scientific naturalism arriving via the explanatory power of computational philosophy, but combined with a generalization of the common ground of having such a perspective.

Just as the classical concept of "Heat" has evolved to discrete models of networked processes involving energy, IE frames classical "Knowledge" into networked models of informational processes. Floridi's elucidations of the new, ontological entropy are quite specific, as well as broadly founded. The infosphere spans four formal classes of properties: Neo-Platonic right to existence, Spinozian right to integrity, a Libertarian right to openness, and a Constructionist right to development. Some modal properties of this entropy are inconsistency of logical possibility; impossibility of implementation; absence of existence or occurrence. Humanistic properties include volatility, transitoriness, ephemerality, instability, loss or destruction, misuse, unauthorized use or modification, improper disclosure, inaccuracy, partiality, in-authenticity, unreliability, poverty, sterility. Illuministic properties are unavailability, secrecy, inaccessibility, and disorder. A Constructionist property of entropy is redundancy. All of these attributes yield from standard philosophical bases and will hopefully be the subject of greater scrutiny and testing by commentators.

Reply to Herold

John Barker
University of Illinois at Springfield

Ken Herold raises a number of interesting points in his response to my paper "Too Much Information" (Barker 2008). However, the fundamental points I raised there remain sound, in my opinion and as I will argue here. After addressing two issues that are basically exegetical, I will argue that the issues I raised in the context of Shannon information theory apply to other notions of information content as well.

First, a note on informational objects as arbitrary objects. Herold takes me to task for interpreting Floridi overly-broadly, and insists, if I understand him correctly, that informational objects constitute only a small subset of all objects. However, I find this hard to square with Floridi's own words. In (Floridi 2008a), he writes:

> In I[nformation] E[thics], the ethical discourse concerns any entity, understood informationally, that is, not only all persons, their cultivation, well-being and social interactions, not only animals, plants and their proper natural life, but also anything that exists, from paintings and books to stars and stones; anything that may or will exist, like future generations; and anything that was but is no more, like our ancestors or old civilizations. Indeed, according to IE, even ideal, intangible or intellectual objects can have a minimal degree of moral value, no matter how humble, and so be entitled to some respect. (Floridi 2008, 12)

It is quite clear from this passage, which is not in any way atypical, that IE treats all objects whatsoever as moral patients. Since IE is also a theory of informational objects as moral patients, it would seem to follow that IE regards all objects as informational objects. Floridi goes on to write:

> IE is impartial and universal because it brings to ultimate completion the process of enlargement of the concept of what may count as a centre of a (no matter how minimal) moral claim, which now includes every instance of being understood informationally (see section 2.4), no matter whether physically implemented or not. (Floridi 2008a, 12)

"Being understood informationally" is ambiguous: it could mean that existence in general is to be understood informationally, i.e., that all objects are informational; or it could meant that IE concerns itself with a special class of entities, namely, informational objects. However, the latter interpretation seems at odds with Floridi's claim that IE is "universal" and that it generalizes other moral theories such as land ethics.

Next, I want to say a few words about the interconnected notions of complexity, entropy, benefit, and harm. In (Barker 2008), I complained that the standard of benefit and harm to informational objects was underdefined. Herold quite rightly points out that the connection between right action and entropy is spelled out reasonably clearly in Floridi's four laws of IE. In short, it seems fair to say that Floridi identifies harm with entropy. Indeed, this was always my assumption, and I accept Herold's correction on this point. I also assumed, as apparently does Herold, that the notions of complexity and entropy are complementary, in that entropy is to be seen as a lack of complexity and vice versa. Thus, if numerical measures of entropy and complexity are given, then the one is simply the additive or multiplicative inverse (say) of the other. I am more than happy to make this assumption, though I am unaware that Floridi ever actually endorses it (though I would be happy to be proven wrong on this).

This brings us to the main topic of this note: namely, whether there is a reasonable notion of complexity (or of entropy) that can underwrite IE. Now there are many different measures of complexity, including the Shannon measure that I discussed in "Too Much Information," as well as various notions of computational complexity. So a natural starting point is to ask: What measure of complexity (or of entropy) does Floridi prefer in connection to IE? Unfortunately, while Floridi has written extensively on information, I am not aware that he has really answered this question. Nor has Herold provided a clear answer, though he does hint that some measure of computational complexity may be appropriate.

In (Floridi and Sanders 1999), Floridi and J.W. Sanders do provide a rigorous framework for talking about entropy as a moral evil. Their account is instructive, if only because it helps illustrate what I am asking for in an account of entropy. Floridi and Sanders define an entropy structure to be a triple $(X, \leq, \sim)$, where

- X is a nonempty set, whose elements are called states;
- $\leq$ is a transitive and reflexive binary relation on X;
- $\sim$ is the equivalence relation on X defined by $x \sim y$ iff $x \leq y$ and $y \leq x$.

An action on an entropy structure is simply a binary relation A on the structure's state space X. Intuitively, x A y means that acting on a state x can produce a state y. If the relation A is many-one (i.e., if it is a function), then A is said to be a deterministic action. Floridi and Sanders then make the following definitions:

- If x A y implies $y \leq x$ for all states x and y, then the action A is entropy decreasing;
- If x A y and $x \leq y$ for at least one pair (x, y), then A is entropy increasing;
- If x A y implies $x \sim y$ for all x and y, then A is entropy invariant.

Floridi and Sanders go on to provide several examples of entropy structures.

The above formalism is the closest thing I have found in Floridi's writings to a formal definition of entropy. However, while the above system may be formal, it is not in any way a definition of entropy. It is rather a formal framework in which a more detailed account of entropy could be developed, or in which several different accounts of entropy could be developed simultaneously. This formal framework simply imposes the following constraint on the notion of entropy: it requires that the relation x is a state of lesser or equal entropy to y be reflexive and transitive. No other aspect of the notion of entropy is thereby explained.

Now I am sure that Floridi and Sanders never intended the entropy structure formalism to be a substantive account of entropy. It is simply a framework in which many such accounts could be given. Indeed, this generality can be seen as a strength, since it abstracts common features of all the various accounts that it can accommodate. That said, if the entropy structure framework were all that could be said in general about entropy, or about the sort of entropy that is relevant to IE, then IE would be severely underspecified. Indeed, I would argue that if all we assume about entropy is that it can be described by an entropy structure, then IE becomes all but empty. It is all but empty (on this assumption) because it imposes virtually no constraint on what is to count as one state being higher in entropy than another, and thus, it imposes virtually no constraint on what is to count as a benefit or a harm to a patient.

To make this clear, consider the following moral theory: any object whatsoever is benefited by bringing it closer to Peoria, Illinois, and harmed by moving it further away. While no one would seriously entertain such a theory, it is easily accommodated by the formal system under discussion. Let our state space be the set of points in space. For x and y in our state space, define $x \leq y$ iff the distance between x and Peoria is no greater than the distance between y and Peoria. Clearly $\leq$ is transitive and reflexive. Defining $\sim$ in the standard way, we see that $(X, \leq, \sim)$ is an entropy structure. Now for any given way of acting on an object O, define the formal analog A of this action to be the relation {(x, y): O was located at x before it was acted on, and is located at y afterward}. Then A is an action on our entropy space. And if we plug all of this into the four laws of IE, what results is a moral theory that tells us to bring as many objects as we can to Peoria.

Thus, if the only constraint we place on entropy is that of entropy structures, then the ridiculous moral theory just described, as well as countless others, fail to be excluded by IE. Now I want to be clear: none of this is in any way a criticism of the entropy structure formalism. My point is simply that in the absence of some further and more specific account of entropy, IE offers little or no recommendation for moral action.

Thus, it behooves us to look for a more substantive account of entropy, or equivalently, of complexity. Numerous such accounts have been given. Taking a cue from Herold, let us now consider the notion of computational complexity. Computational complexity is typically defined as a measure of the complexity of a decision problem, such as that of deciding whether a string s over a finite alphabet $\sum$ belongs to a given set X. Glossing over several details, the complexity of the problem

may be defined in terms of the minimum number of steps that it would take a Turing machine to decide the problem. Along these lines, Herold mentions the interesting work of Gregory Chaitin. In (Chaitin 1990), he defines a complexity measure for strings, not decision problems. Essentially, the complexity of a string s is defined to be the size of the shortest program that produces s as its output.

Now I have no quarrel with the value or technical feasibility of this or any other formal measure of complexity. My worry concerns its application to concrete cases. Agents act on physical objects (including other agents). If there is a moral imperative to reduce entropy and increase complexity—let alone if this is the fundamental moral imperative—then some complexity measure for concrete, physical objects must be given. Now the complexity measure we want may seem pretty obvious. Let H be some complexity measure for strings from some fixed alphabet—H might be the Chaitin measure, for example—so that H(s) is s's complexity. Now let t be some physical object; we want to assign a complexity to t. And the natural way to do this is as follows: define H(t) = H(s), where t is a token of s. Simple and obvious though this move may be, however, it raises some important issues.

What is it, in general, for a physical object t to be a token of a string s? What, for example, does it take for a physical object t to be a token of the string "hat"? The answer has very little to do with the physical properties of t. Instead, it is a question that we settle by convention. Questions about tokens of "hat" don't arise, and are not intelligible, unless there is some convention that determines which objects token which strings. There is simply no sensible question of what string a given physical object tokens qua physical object.

Moreover, different conventions will result in assigning different complexities to any given object. To see this, fix a finite alphabet $\sum$, and let F be a 1-1 function from $\sum^*$ onto $\sum^*$, where $\sum^*$ is the set of all finite strings from $\sum$. We will assume that F is computable, but beyond that we will make no assumptions about F. Now let us define a new convention regarding types and tokens: if an object t is a token of a string s under the standard convention (whatever that may be), then t is a token of F(s) under our new convention. Obviously, nothing prevents us from adopting this new convention. In fact, we adopt conventions like this all the time: it is called encryption. But it is also obvious that F need not be complexity preserving: H(s) need not equal H(F(s)).

Thus, it appears that the complexity of a physical object, on this approach, is partly a matter of convention. There is no such thing as the Chaitin complexity of a physical object qua physical object: to determine an object's complexity, one must first decide what string it tokens. At this point, a few observations are in order. First, this result does not depend in any way on the details of Chaitin's definition of complexity. It applies to any measure of complexity defined on strings of symbols. Second, what we have found here seems to me to be a very general phenomenon. As I argued in "Too Much Information," if we use the Shannon measure of complexity in place of the Chaitin measure, then the complexity of a physical system depends on our choice of a probability measure p—indeed, it is simply $\log_2(1/p)$—and it is not obvious that anything constrains our choice of p. In either case, we have a mathematically well-defined notion of complexity, but to apply that notion to a physical object, we must first subsume that physical object under the mathematical formalism to which the complexity notion applies in the first instance; and in general there is no privileged way of doing this.

Now none of this is a problem for any of these complexity notions if they are used for their intended purpose. Complexity measures are used to measure information, and when we apply these measures in a real-world context, we have already adopted the conventions that allow us to speak of physical systems as information and not just as matter. But when we turn complexity into a morally significant quantity, something we have a moral reason to try to maximize, then things become more complicated. If the complexity or entropy of a physical system is (partly) a matter of convention, then so is the moral worth of that system, and therefore so are the moral claims that it exerts on us. This issue becomes particularly acute if information ethics is regarded as general ethics, as Floridi seems to claim, for then the convention-relativity of information ethics infects all of ethics.

This point about relativity to a convention was one of the primary themes of "Too Much Information," and I hope I have made it a little clearer here. I would like to conclude by addressing Herold's point about the method of Levels of Abstraction (LoAs). Namely, he suggests that the problems I raise would go away if I simply applied that method, though he does not elaborate on this thought. On the contrary, I would argue that the method of LoAs simply throw the problems I have been discussing into sharp relief.

LoAs are perhaps best explained by using Floridi's example of a traffic light. A typical American traffic light has three states: red, yellow, and green. In saying this, we are of course abstracting away from all the myriad physical details of traffic lights. We are adopting an abstract model of traffic lights, a model that includes the three abstract states red, yellow, and green. Much of what we have to say about traffic lights can (and should) be stated solely in terms of these abstract states, without worrying about the physical states that realize them. We have just described traffic lights at a certain level of abstraction. If we like, we can describe traffic lights at a different level of abstraction, one that includes more, or fewer, or different, details. For the exact formal details, see, for example, (Floridi 1998b).

LoAs are ideally suited to describing information. Consider our traffic light example again. In describing the traffic light in terms of the states red, yellow, and green, we are effectively describing it as a system that contains $\log_2(3)$ bits of information. Or think of how we describe computers. We can describe a computer as a piece of physical stuff if we like; but when we describe a computer as a computer, we generally describe it as realizing some abstract machine or other. That is, we describe it at a higher level of abstraction than the physical level.

However, there is nothing in all this talk of LoAs that actually solves the issues I have raised above. LoAs simply provide a convenient language for discussing them. Indeed, the information content of a physical object is, pretty clearly, relative to the LoA at which we choose to describe it; and we are free to adopt whatever LoA we please in so doing. Consider again the problem of applying the Chaitin complexity measure H to a physical system. The problem was that before we could do this, we had to first define the type-token relation, associating physical states t with strings s in $\sum^*$. Now we can think of this choice of a type-token relation as the adoption of a LoA, one that describes its objects as (tokens of) strings and abstracts away from low-level physical details. However, nothing prevents us from adopting an alternative LoA, so that an object that is identified with a string s in the first LoA is identified with the string F(s) in the second LoA. This flexibility in the choice of a LoA is a fundamental feature of the method of LoAs. It is also, of course, precisely the feature that, I am arguing, makes informational complexity a questionable measure of moral worth.

References

Barker, John. 2008. Too much information: Questioning information ethics. APA Newsletter on Philosophy and Computers 08(1).

Chaitin, Gregory J. 1990. Algorithmic Information Theory. Cambridge: Cambridge University Press.

Floridi, Luciano and J.W. Sanders. 1999. Entropy as evil in information ethics. Etica & Politica, special issue on Computer Ethics, 1.2. Oxford University, Computing Laboratory, Programming Research Group Technical Report TR-5-00.

Floridi, Luciano. 2008a. Information ethics, its nature and scope. In Moral Philosophy and Information Technology, edited by Jeroen van den Hoven and John Weckert. Cambridge: Cambridge University Press.

—. 2008b. A defence of informational structural realism. Synthese 161(2): 219-53.

DISCUSSION 3: ON LOPES

CO-ORGANIZED WITH THE AMERICAN SOCIETY FOR AESTHETICS (ASA)

Videogames, Interactivity, and Art

Grant Tavinor
Lincoln University in New Zealand

Videogames are one of the most significant developments in the popular arts in the last fifty years, and a great deal philosophical interest arises from their artistic employment of computer technology. Whereas other artistic media such as music and film have felt the effects of digital technology—especially concerning changes in digital modes of production and distribution—the development of videogames seems to constitute the growth of an entirely new artistic medium (Tavinor 2009).

Videogames and Interactivity

Insofar as the technological and formal developments make videogames distinctive as an artistic medium, this distinctiveness arises from their interactivity. There is an immediately plausible sense in which videogames are more interactive than some traditional artistic media: unlike the movie Star Wars where the viewer is passive in respect to the events on the screen, in the videogame Star Wars: The Force Unleashed the player may adopt the role of a character such as Darth Vader through which they can act in the fictional world of the game.

There are doubts, however, about the usefulness or accuracy of describing videogames as interactive. Games theorist Espen Aarseth opposes applying the term to videogames, thinking the term "interactive fiction" either meaningless or trivial (1997, 50). Indeed, there may be some warrant for the charge of triviality: videogames are games, and calling a game "interactive" seems redundant. A further difficulty is that maintaining that videogames are interactive, implies, somewhat problematically, that traditional media are "passive" in some respect: but all art is interactive insofar as it involves appreciators in a physical and cognitive engagement with a work. Finally, as Dominic Lopes notes, because of its sheer ubiquity as a technological buzzword, the concept of "interactivity" is prone to abuse, and is of limited theoretical use without specifying a substantive meaning (2001, 67).

Indeed, we can frame the interactivity of videogames by drawing on Lopes' theory of digital art. Lopes argues that a number of recent artworks, exploiting the representational potential of computers, allow appreciators modes of interactive engagement that "no other art media can enjoy" (2003, 112). Lopes' theory, developed to address digital artworks, promises to apply to videogames because he sees traditional game activities as a paradigm of the kind of interactivity now seen in digital art. Distinguishing between "strongly interactive" works and "weakly interactive" ones, he claims that

> Games are "strongly interactive" because their users' inputs help determine the subsequent state of play. Whereas in weakly interactive media the user's input determines which structure is accessed or the sequence in which it is accessed, in strongly interactive media we may say that the structure itself is shaped in part by the interactor's choices. Thus strongly interactive works are those who structural properties are partly determined by the interactor's actions. (2001, 68)

Much of what is referred to as interactive in the digital realm, is, he concludes, only weakly interactive because it involves an appreciator merely navigating their way through a predefined structure. Games like chess, however, are strongly interactive because the sequence of game states is determined by decisions made by players given the starting state and the rule set or "algorithm" that defines the permissible state transformations or moves (Juul 2005). This characterisation of the strong interactivity of games can be applied in the case of many interactive artworks because they share a productive algorithmic structure with games. When the interactive object in question is an artwork, the structures in question are those that are behind "whatever intrinsic or representational properties it has the apprehension of which are necessary for aesthetic engagement with it" (2001, 68).

It seems clear enough that videogames do count as strongly interactive in Lopes' sense: videogames do not merely involve choosing the order in which the representational structures of the game are experienced, but involve the player having an effect on just which potential structures of the game are depicted, and how those structures are depicted. Playing as Niko Bellic in Grand Theft Auto IV, the player does not merely cue the representation of parts of an artwork that have been previously encoded, as they might by choosing in which order to read the chapters of a novel or listen to tracks on an album—both of which are among Lopes' examples of weak interactivity (2001, 68-69). Rather, players shape what actually occurs in the game. My playing of the Grand Theft Auto IV is likely to be unique to me in that the fictional events that occurred in my playing of the game were dependent on my decisions: the game in all its detail was rendered only after I had my input.

Thus interactivity is tied up with the ontological issue of videogames as multiple instance works, because videogames seem to be work types that necessitate the interaction of a player before they are instanced as tokens. In setting out his ontology of mass artworks, Noël Carroll notes that though a type/token distinction is crucial for capturing something of the relationship between multiple instance work types and their instances, the logical distinction is not "fine grained enough" to capture what it is that instances various forms of type artworks (1998, 212). A theatre performance, he claims, is instanced by an "interpretation" of a script; a film is instanced by the screening of a "template." Videogames are clear cases of multiple instance works, and there are equally clear differences to other kinds of art in the way they are instanced. The representational artefact at the basis of a videogame like Grand Theft Auto IV is not a template from which the appreciated work is shown, or a script that is interpreted. Rather, the kind of interaction that is crucial in instancing a videogame is a playing, whereby the player "reveals" something of the structure of the game through their interaction (Lopes 2001, 74). The work relies on an algorithm

that makes possible any number of varied renderings, which depend for their detail on the input of the player.

The structures that comprise an instance of a videogame are various kinds of graphical, aural, textual, and even tactile representations on a display device, because it is these things that a playing has the role of rendering from the game's algorithm. Furthermore, given that these representations are almost always of fictional events, in depicting situations with an imagined existence only, we might in a Waltonian sense say that the structure that is being determined by the interactor's choices is a perceptually modal "prop" that depicts a fictional world (Walton 1990, 21). Hence, videogames may be "interactive fictions" in a theoretically robust sense (Tavinor 2005).

Effectively, the control of fictive events that in film is invested in the choices of the director, writer, editor—because it is they who play the crucial role in encoding the template from which the film is shown—is ceded somewhat to the player. Whereas in the case of film the audience encounters the work after it has been rendered in the form of a film reel or digital file, in videogames the imaginative prop is rendered only after the player interacts with it, and in a way that accords with their own imaginative and participative activities regarding the prop. This means that whereas previous audiences were somewhat passive in respect to what was rendered by the work, players in videogames genuinely are active contributors to the fictions and narratives of the games they play.

Of course, this interactivity derives squarely from the technology of computers, which, through their algorithmic structures, act as representational props able to render audiovisual representations in real-time, contingent on the input of the player. The means in which videogames have achieved these representational ends, involving things like game-engines, polygonal models, virtual cameras, the graphics pipeline, and so on, are immensely interesting in their own right (Tavinor 2009, 61-74).

Videogames as Interactive Art

I have claimed, without anything in the way of argument, that videogames are art works; and moreover that they are interactive art. Two difficulties loom given these claims. First, there are the general worries with the art status of videogames. Second, there is a worry that even if videogames are art, that their artistic aspects are not in fact interactive ones. Remembering Lopes' claim that traditional gaming is a paradigm of interactivity, it might be argued that even though some videogames are properly called art, they are interactive only in virtue of their nature as games. Indeed, it seems that is just the case with a great many videogames where key artistic structures—such as narratives—lack the interactivity characteristic of the gameplay. In a game such as Metal Gear Solid 2, the gameplay intermittently pauses so that the game's story can be conveyed by short, pre-rendered films. Furthermore, the story that is told by these cut-scenes is identical for all playings irrespective of what the player does during gameplay. In such cases, the art of the game may seem to constitute a non-interactive or merely weakly interactive artistic veneer on a strongly interactive game.

Given the space I have here I will not address the first issue directly and argue that videogames are indeed art; this is an issue that has been discussed elsewhere (Smuts 2005; Tavinor 2009). But I will address the second issue, and claim that the structures that are determined by user interaction in many recent artistically inclined videogames are the representational states that are crucial to both the gaming aspects and artistic aspects of the videogame. The reason for this is that increasingly these two aspects reside in a single representational structure. What then is interactive about videogames qua art?

Because recent videogames depict their games in representationally rich ways, the principal strongly interactive artistic structure in many videogames is itself the gameplay. Gaming has been a major part of the Western conception of the arts, and though some games may have aesthetic properties—a gambit in chess might be described as elegant—this has not frequently been the basis for calling games like chess art (but see Humble 1993 and Osborne 1964). But the games to be found in videogames often do seem to be depicted in an artistic way because of their representational nature as complex fictions. The moves and objectives in many recent videogames are not mere formal possibilities with little representational significance, as they might be in a game like chess or checkers, but stories of survival depicted through aesthetically engaging fictional worlds.

For example, the rules and objectives in the post-apocalyptic role-playing shooter Fallout 3 are defined by the fictional abilities of the player-character and their fictional goals, and the game is about surviving and advancing in the gameworld of the Capital Wasteland. To do this the player must battle the adversaries they find in that world, scrounge for resources, and interact with the gameworld inhabitants through conversation and other (often more violent) means. Because of the genuine artistry of the game, many of the gameworld encounters have an extraordinary sense of atmosphere and style: emerging from their fallout shelter at the beginning of that game, the player is struck by the glaring bleakness of the post-apocalyptic world; or, deep in the Wasteland, close to nightfall, the player encounters the Dunwich Building, and the tale of horror within. Gamers and game critics also describe gameplay in aesthetic terms, and evaluate it in ways strikingly similar to the evaluative practices of traditional art audiences: best evidence for this are the reviews and critical pieces to be found in the growing gaming literature.

Many recent games, especially of the "sandbox" or "open-world" variety, encourage gameplay in the form of a free aesthetic exploration of a fictional world. The exploration of an aesthetic environment, such as Liberty City in Grand Theft Auto IV, or the fantasy province of Cyrodiil in The Elder Scrolls: Oblivion, is strongly interactive because though the graphical environment is based on a determinate 3D model, the artistic structure that is ultimately rendered depends on the explorative activity of the player. Technically, this "exploration" is determined by the player's control of the virtual camera that is used in 3D games to define the player's fictive perspective on the gameworld (Tavinor 2009, 67). Open-world games are often played with aesthetic motivations, with the player framing the virtual world in an aesthetic way. The games themselves encourage these kinds of aesthetic playings: Grand Theft Auto IV gives the player access to a helicopter, and one of its most alluring uses is to take scenic flights to experience the significant dynamic beauty of Liberty City. The Elder Scrolls: Oblivion contains hard to access locations in the mountains that seem placed there solely to encourage an aesthetic exploration of the environment.

A further key reason why it is tempting to characterise gameplay as artistic is because of its rich emotional nature: increasingly, gameplay has the ability to depict rich first-hand fictional experiences that draw on the player's emotions. In videogames, the player can have kinds of emotions that depend on their ability to be a participant in an emotionally provocative situation: it is possible to be worried about harming a fictional character, guilty for having done so, or even have feelings of sympathy and care for the characters in, as the game BioShock demonstrates. Hence videogames have the ability to involve the player's emotions in a way that may be denied by traditional

non-interactive fictions such as novels and films. Again, this is because the game—the rules and objectives—is presented as a set of challenges and obstacles in a fictional world.

Ultimately, that both the gaming and art of such games are generated by their interactive fictions, means that art and gaming coincides in the single structure of an interactive fiction. The art of such games is not a mere gloss on the game (as it may very well have been in earlier instances) but is the means through which the game is depicted.

The interactivity of videogames has an impact on how they are evaluate for their artistic qualities, because gauging the artistic qualities of videogames demands repeated playings. Lopes argues that in strongly interactive digital art, "the contours of the work type are drawn by what interactions it makes possible" (2003, 112). Similarly, the range of playings made possible by a videogame discloses the true extent of its artistic properties. Getting a real sense of the achievement of a sandbox game such as Fallout 3 demands that the player approach the game on a number of occasions and in differing ways, and indeed, that game supports some very different playings. This, of course, explains something of the immense depth and replay value of such open-world games: whereas a linear first-person shooter may be finished in ten or so hours, open-world games can support hundreds of hours of gameplay.

The formal developments in videogaming rest on the empowerment of the player as an actor having a substantive interaction with the artistic prop whereby something is revealed of the digital artwork. This interactive potential cannot be altogether irrelevant to the question of videogaming's rapid growth and popularity in recent times, and it may be that this popularity is in part based on the fact that videogames do engage players in a more richly interactive way than other fictions. Though videogames may still be in a state of artistic adolescence, there is surely already enough evidence that their interactivity is not lacking in artistic promise, and that they do have the means to be a genuinely valuable and distinctive form of art.

References

Aarseth, E. 1997. Cybertext: Perspectives on Ergodic Literature. Baltimore: Johns Hopkins University Press.

Carroll, N. 1998. A Philosophy of Mass Art. Oxford: Clarendon Press.

Humble, P. N. 1993. "Chess as an Art Form." British Journal of Aesthetics 33(1).

Juul, J. 2005. Half-Real: Videogames Between Real Rules and Fictional Worlds. Cambridge, Mass.: MIT Press.

Lopes, D. M. 2001. "The Ontology of Interactive Art." Journal of Aesthetic Education 35(4).

Lopes, D. M. 2003. "Digital Art." In The Blackwell Guide to the Philosophy of Computing and Information, edited by Luciano Floridi. Oxford: Blackwell.

Osborne, H. 1964. "Notes on the Aesthetics of Chess and the Concept of Intellectual Beauty." British Journal of Aesthetics 4(2).

Smuts, A. 2005. "Are Videogames Art?" Contemporary Aesthetics 3.

Tavinor, G. 2005. "Videogames and Interactive Fiction." Philosophy and Literature 29(1).

Tavinor, G. 2009. The Art of Videogames. Oxford: Wiley-Blackwell.

Walton, K. 1990. Mimesis as Make-Believe. Cambridge MA: Harvard University Press.

Online Education

Gender and Online Education

Margaret A. Crouch
Eastern Michigan University

Introduction

At the beginning, some thought that the Internet would make differences in gender, class, race, age, and disability irrelevant. This forecast was embodied in the 1997 MCI commercial called "Anthem." The voiceover said:

> People can communicate mind to mind.
> There is no race.
> There are no genders.
> There is no age.
> There are no infirmities.
> There are only minds, only minds.
> Utopia?
> No.
> The Internet.
> Where minds, doors, and lives open up.
> Is this a great time or what? (Goldman, Papson, and Kersey 2003)

Some researchers call this the "democratic" theory or model of computer-mediated communication (CMC). It was thought that the absence of the social cues that usually occur in face-to-face interactions would decrease the use of stereotypes and other forms of domination or social exclusion (Yates 2001, 22). And, yet, we are all here today to discuss the falsity of this prediction. The exultation with which this possibility was heralded implies that differences in gender, class, and race, among others—what we now refer to as "diversity"—are negative; which, interestingly, conflicts with the most recent arguments for the value of diversity in education. I will have a bit more to say about this at the end of the presentation.

I have been teaching an online critical thinking course for about ten years. All of the philosophy faculty at my institution teach online courses. It seems we are not alone.

According to the National Center for Educational Statistics (NCES) 2006-07 survey, the number of online courses offered by degree granting postsecondary institutions, and the number of students enrolled in online courses, is steadily increasing (Parsad and Lewis 2008). In 2006, 61% of two- and four-year institutions reported offering online courses, compared to 56% in 2000. In 2006, there were about 9.4 million students enrolled in online courses compared to just over 3 million in 2000. In 2005, women made up 57.4% of all enrolled undergraduates in two- and four-year institutions (National Center for Education Statistics 2007). Women also make up the majority of students enrolled in online courses.

I do not have any statistics on the number of philosophy courses online, or the number of philosophy programs offering courses online, or even the number of degrees in philosophy offered online. There is very little written on good ways of designing online philosophy courses.[1] So, most of what I discuss here will not have direct bearing on online philosophy education, though it certainly has an indirect bearing on it.

There are many different terms used for what I am calling "online education." Some of the material I will discuss refers to courses that are offered entirely online, others to so-called hybrid courses. Other terms for online courses include asynchronous

learning environment (ALN), computer-supported learning (CSL) environments, and eLearning. Much of the research on gender and online education is on the aspect of online courses called "threaded discussions," which are a kind of asynchronous, computer-mediated communication (CMC).

I will come at the topic of gender and online education from three different angles. I will start with the question what differences, if any, male and female students experience in online classes. I will then move to the question of how accessible online education is to women in the United States and other developed countries. I will end with the question of the potential of online education to promote women's equality around the world.

1. What differences, if any, are there in the experiences of male and female students in online classes?

There are many aspects to this question, but I will focus on two: online discussion and learning style. Studies of face-to-face classrooms have noted that men tend to dominate discussion in mixed-gender classes. This makes such classrooms less welcoming to female students than they might be. Does this transfer to online discussions? Or, does the relative anonymity of online discussion, where discussants cannot see one another, mitigate this dynamic? (Hamann, Pollock, and Wilson 2003)

The data on gender communication in online courses is problematic. Most of the studies are based on very few students, often in education or technology courses, many of them graduate courses. This makes generalization to fifty-person undergraduate philosophy courses, such as the one I teach, questionable. First, the samples are very small. Secondly, the subject matter of the courses often might be classified as "training" rather than education in the sense we in the liberal arts understand it. And, thirdly, graduate courses involve different sorts of students than undergraduate service courses. So, I will discuss some of the data from studies on gender and online classes, but we should be careful about generalizing. This probably won't matter that much, however, since there is so much disagreement among the studies. Some argue that women are still at a disadvantage in online courses, some that they are advantaged, and some that they are neither.

Much of the research on gender and online education is based on the analysis of asynchronous online text discussions. Such discussions are an important part of online courses, and provide a convenient set of data for researchers. As I said above, one of the issues on which researchers have focused is whether the gendered forms of communication that take place in face-to-face classrooms are similar in online course discussions. From my reading of the literature, I would say that the answer to this question is, maybe.

Some scholars argue that gender differences appear in online discussions pretty much as they do in face-to-face discussions. For example, the much-cited Herring studied computer-mediated communication (CMC) in the form of public ListServs and other discussion forums in the 1990s and found many gender differences. In a summary article published in 2000, she says of this early research (and that by others):

> In asynchronous CMC such as takes place in discussion lists and newsgroups on the Internet and Usenet, males are more likely to post longer messages, begin and close discussions in mixed-sex groups, assert opinions strongly as "facts," use crude language (including insults and profanity), and in general, manifest an adversarial orientation towards their interlocutors (references omitted). In contrast, females qualify and justify their assertions, apologize, express support of others, and in general, manifest an "aligned" orientation towards their interlocutors. (References omitted, Herring 2000)

There is much more. It does seem pretty well established that the relative anonymity of computer-mediated communication does not eliminate gender entirely because of the gendered styles of communication people tend to use.

Yates, who is frequently cited in support of the view that women are disadvantaged in online education, also argues that computer-mediated communication (CMC) incorporates many of the same sorts of inequalities found in face-to-face discussions, thereby marginalizing the same groups that are marginalized in face-to-face interactions (Yates 2001). In a recent survey article, Yates goes over the findings of Herring and others (a circular sort of support) and claims,

> These conclusions would seem to have quite negative implications for use of CMC in education; especially as the material studied by Herring consisted of academic and educationally-orientated CMC interactions. The "democratic" model has not won out and, as with face-to-face educational situations, gender has a key role to play in structuring the interactions so as to marginalise women's contributions. (Yates 2001, 27)

Yates and Herring are both somewhat optimistic, however, since they do not think that CMC is inherently sexist. They believe that the sexist dimensions of online communication can be reduced by changing social context. For example, Herring says that

> Some evidence suggests that women participate more actively and enjoy greater influence in environments where the norms of interaction are controlled by an individual or individuals entrusted with maintaining order and focus in the group. …Female students also participate more—sometimes more then male students—in online classrooms in which the teacher controls the interaction, even when the teacher is male. (References omitted, Herring 2000)

Most of the research on which Herring and Yates base their conclusions was performed in the early to middle 1990s. More recent research both confirms and disconfirms the findings of Herring and Yates.

An example of a study confirming the claim that males and females communicate differently in online discussions is a recent study by Gougeon (1998). He examined the "communication techniques" of fifteen women and four men enrolled in an online course. Gougeon used Debra Tannen's "framework for interpersonal communication" in analyzing the discussions. This posits that men and women have different needs and so have different "purposes" in communication and fulfill those purposes using different sorts of conversational gambits. Men seek to establish status first and foremost, while women seek to establish relationships. Establishing status requires creating distance from others and emphasizing differences from others. Establishing connection requires emphasizing similarities with others. In his analysis, Gougeon found that the men and women posted different kinds of comments, with the men most concerned with establishing their status to the group, and the women more concerned with establishing relationships. The men "reported" to the discussion, while the women "developed patterns of communication supporting the following: a sense of intimacy among participants, equal or horizontal alignment in status, symmetry based on the establishment of similar experiences, and a sense of interdependency with other group members."

This is a fairly typical study of its type, using Tannen's sociolinguistic analysis of communication practices. Such studies usually find different kinds of communication in online discussions depending on gender.[2] However, and most of these researchers would agree, there are many possible reasons for these differences, and ways of challenging them through the design of discussion fora.

For example, a study by Bostock and Lizhu (2005) found that women and men posted different numbers of messages depending on the gender composition of the group, though all the messages were equally "cognitive" (as opposed so "social"). All female groups posted more messages per student than mixed or all-male groups. Males posted more messages in mixed-gender groups than in all-male groups, but females posted fewer messages in mixed-gender groups than in all-female groups. Interestingly, they also found that the women in the courses being studied tended to prefer online discussions to face-to-face discussion and to get higher grades in the courses.[3]

Some argue that far from being disadvantaged or marginalized in online discussion, women are actually advantaged. Coldwell et al. conducted a large scale survey of students in New Zealand regarding both their perceptions of online education and their willingness to express themselves in discussion in online courses. They found "no significant differences between female and male students with respect to being able to use the online learning environment confidently and effectively. In general, female students were more willing to participate in online discussions."[4]

Other studies seem to confirm the view that women prefer online discussion to face-to-face classroom discussion. A study comparing online and face-to-face discussions by Caspi, Chajut, and Kelly (2008) found that "men over-proportionately spoke in the face-to-face classroom whereas women over-proportionately posted messages in the web-based conference" (718). They cite studies supporting the view that men dominate face-to-face classroom discussion and conclude that, "while in the classroom, women may be deterred from active participation because of an expected, imagined or actual threatening climate, in WBIE [web based instructional environment] they may feel less intimidated" (723).

This suggestion is supported by a study of female experiences of online courses by Patrick Sullivan (2002). Female students said that they liked the relative anonymity of online courses. I found the actual responses very interesting and include a few of them:

- One positive point for the students is that we do not have to face each other in a classroom atmosphere and be intimidated by looks, weight, height or personalities. Being online gives everyone the same advantage…
- In an Internet course we are unable to judge people by appearance, we have no idea what the other students look like, or what their ethnicity is. I think that is one thing that makes Internet classes so great. Some of the students in my class this semester, in my opinion, are brilliant writers, and I could tell you if they are male or female, but almost no other physical characteristics about them. There is no stereotyping or bias amongst the students here and no opportunity for bias by a teacher.
- It's easier to be yourself if you're "invisible." …When speaking in the traditional classroom, everyone's attention is focused on you in unison—literally all eyes are on you, and if someone disapproves or disagrees, it's obvious (body language, roll of the eyes, etc.). I think the reason why people don't participate more in classroom discussion is because they are afraid of looking dumb or being judged in some way. The anonymity of the online classroom removes those fears completely because you don't know your classmates' reactions to what you said for a few days.
- The type of asynchronous bulletin boards that most classes use allow for thoughtful responses. They take away the need to be first with your hand up, to feel like you have to think very quickly on your feet. They take away the need to deal immediately with someone's response to you—again allowing for time to think. (138-140)

My survey of literature on gender and discussion in online courses leads me to a tentative conclusion. The gendered social context of life outside the online course makes its way inside, so that gender still figures in online discussions. However, women seem to be holding their own and many indicate that they like online discussion more than face-to-face discussion. It may be that they experience online discussion as less discriminatory than face-to-face classroom discussion.

There is much less research on "learning styles" in online education, though it poses an interesting question. Is the online learning environment alien and forbidding to women? Despite the worries, this seems not to be the case. There are ways of rendering the online educational environment very congenial to women learners.

Pamela Whitehouse (2002) argues that online learning can be made very welcoming for female students depending on course design. Taking seriously the Belenky et al. claim that women learn differently than men, Whitehouse developed what she calls the "feminist distributed learning model." This model "builds a community of active learners" and makes use of different methods and technologies to reach different sorts of learners. The important aspects of the model are that students construct knowledge together and individually, they take responsibility for their own learning, that students feel supported by other students and the instructor, and that there are "multiple entry points for different kinds of learners" (Whitehouse 2002, 216-17). Examples of this latter are different forms of communication. For example, the platform I use has chat rooms for synchronous communication, threaded discussions for asynchronous communication, and e-mail for personal communication with other students or the instructor. Whitehouse says that the threaded discussion, in particular, is conducive to learning for women students. No one can dominate the discussion, and shy students can take the time they need to contribute with less anxiety. Further, the instructor's facilitation of discussion forums can help to create safe spaces where students can ask any question, no matter how elementary, about the course, the technology, etc., but also share their views and experiences with other students. A further result of such a course is that students learn to be comfortable with the technological environment, using skills that are necessary for the contemporary workplace. As Whitehouse says, "Students learned to use materials posted online, develop their own online material, share their thoughts and their work effectively, and make decisions using the most appropriate form of communication for a given situation" (2002, 221). This is an important part of Whitehouse's argument about the feminist nature of online or blended teaching and learning: "what we can gain from modeling how new technologies can be used in ways that make the culture of computing accessible to women" (2002, 222).

A 2002 report by Kirsten Montieth, co-ordinator of the Centre for Research and Development in Learning Technology at University of Stirling on online learning styles based on a content analysis of online discussions contends that "the distinction between traditionally male and female learning styles has become blurred. While male students are retaining elements of a separate learning style they are shifting towards a more connected learning approach, traditionally associated with female learners" (Monteith 2002, 2). She interprets this to mean that "the virtual classroom is becoming a female domain" (Monteith 2002, 2).[5]

2. How accessible is online education to women in the U.S. and developed countries?

In the 1990s, there were reports of a "digital divide" between the genders, with males having more access to computers and much greater competence with computers. In the United States and other developed countries, this divide seems to have disappeared.

According to the U.S. Census Bureau, as of 2003, 61.8% of U.S. households reported the presence of a computer, and 54.7% had Internet access. Broken down according to gender, 65.7% of men and 57.4% of women have computers in their homes; and 58.7% of men and 50.1% of women have access to the Internet from home. Broken down by race, 63.9% of Whites, 44.6% of Blacks, 72.9% of Asians, and 44.3% of Hispanics report having computers at home, while those with access to the Internet from home are: White 57%, Black 36 % Asian 66.7%, Hispanic 36%. Both having computers at home and access to Internet from home track directly with level of education and level of household income.

The Pew Internet and American Life Project December 2008 survey reports that 75% of women and 73% of men use the Internet. The age group with the highest percentage of users is 18-29, with 87%, but 30-49 year olds are close behind with 82%. Seventy-seven percent of whites, 64% of blacks, and 58% of Hispanics use the Internet. Internet use tracks income and education level. The latter statistics are stark: only 35% of those with less than a high school education use the Internet, while 85% of those with some college use the Internet.

Of those using the Internet that are 18 or older, only 6.6% reported taking an online course, 6.3% of males compared to 6.8% of females. The highest percentage of people taking online courses is those of traditional college age. With regard to race, 6.4% taking online courses are white, 6.9% Black, 8.9% Asian, and 6.1% Hispanic. These numbers did not track directly with household income (U.S. Census Bureau 2003).

Distance education has long been a means for women to achieve education. In the late nineteenth century, the Society to Encourage Studies at Home was developed by Anne Elliott Ticknor. Ticknor was the daughter of a Harvard professor. Her society provided educational materials and personal correspondence from a tutor through the post. Students could study history, French, English, science, German, or art at their own pace. Prominent women, such as Alice James, were recruited to serve as correspondents for those enrolled. At once time there were as many as 7,000 students in the society. One of Ticknor's collaborators on the project, Elizabeth Cary Agassiz, referred to the society as "the silent university," and said that "it was intended to change women's lives without altering or impairing the role society had sanctioned for them."[6]

It is striking how little has changed. Many women, both in the United States and in other countries, still seek to fit their education into lives dominated by caring for a family and home, and sometimes also by outside work. In many ways, the availability of online education, the most recent form of distance education, still allows women to "change their lives without altering or impairing the role society has sanctioned for them." This is a theme that we see repeated in study after study.

One of the main reasons that students give for enrolling in an online class is "convenience." Convenience for residential students can mean that they don't have to go to class at a specific time each week and can work in their jammies. But, for many, especially adult learners, online education is their only opportunity for furthering their education. For example, in a Cypriot graduate program delivered entirely online, "the only opportunity they had for graduate education because of its convenience"(Vryonides and Zembylas 2008). However, this "convenience" can be undermined when students realize that they cannot just slot study and coursework into days full of family and work obligations. In a way, the hope of being able to further one's education without disrupting the rest of one's life is a vain hope. A woman in the Greek study said it well:

> For us [women students] time for studying is a luxury…Everything else needs to come first and we have to wait until everybody goes to bed…for things to settle at home in terms of noise and then to switch on the computer. It is funny but when I talk with my male fellow-students they never seem to be mentioning things such as home, children, housework…whereas for us it is a common topic of discussion. (Secondary school teacher, age 40) (Vryonides and Zembylas 2008)

Another woman in this same study expressed a theme common to women in the United States and Greece:

> Even though my children appear to be happy with what I am doing, on one occasion one said to me "Mama, you love your computer more that you love us…" There are moments when I get filled with guilt because my children need me and I wonder if I am stealing the time that is rightfully theirs. (Primary school teacher, age 40) (Vryonides and Zembylas 2008)

This echoes adult female students from decades past, who were not talking about distance education, but face-to-face education. In other words, many of the same barriers remain.[7] The Cypriot women expressed similar stresses, and all of them said that they had to study late at night, after all of their other responsibilities were fulfilled. One connected this to the potential of distance education to provide opportunities to adult learners:

> The open and distance learning program ends up being a painful and exhausting process for someone who works, and especially for those who have family and professional responsibilities. This shows how difficult it is to put an end to social and educational inequalities […]. On the one hand, I was given an opportunity to study, one that I did not have in the past, so I truly appreciate this. On the other hand, however, I cannot benefit from this and so I am deeply disappointed. I have so many responsibilities on my shoulders (family, professional, and social) and the demands of this programs are unrealistic, in my view. So, I wonder: To whom is this program really addressed? If you want my opinion, I don't think it is addressed to women professionals. (Primary school teacher, 38 years old) (Vryonides and Zembylas 2008)

In 2001, the AAUW published a report on women and online education entitled "The Third Shift" (Kramarae 2001). The title comes from the author's claim that, for many women students, education is a "third shift" in addition to paid work

and work at home. Distance education has allowed women to fit further education into their lives, and, in fact, in 2001, the average distance education student was thirty-four, had part-time employment, had some college credit, and was female (Kramarae 2001, 4). This has not changed (Parsad and Lewis 2008).

The most significant obstacle to online education is cost. Tuition for online courses is usually the same as for face-to-face courses, and some forms of financial aid won't support online coursework. Computer-related costs are another barrier. Some women with families mentioned that they had to "fight for time" on the family computer in order to complete their course work. In addition, students are expected to know how to use a computer and to have reliable Internet access.

As we found in the Greek survey mentioned above, one of the most prominent themes in the AAUW study was that women with children found online learning provided them an opportunity that they would not otherwise have had to further their education. Such women have little time for anything but work and family, and online courses enable them to fit study in when they can…after 11 p.m., for example. Both women and men face time pressures, but both women and men said that women, in particular face time pressures because of the responsibilities for care-giving.

> Many women indicate that the third shift of student life occurs late at night or early in the morning. While distance learning allows women to squeeze their studies around the seemingly immovable barriers of family and work life, this evades any general social discussion of how time and responsibilities, both in the work force and the home, might be reconfigured to make fulfillment of educational goals a more humane and less taxing process. Instead, women make individual compromises and choices—as family members, workers, and students—to fit all of these activities into short days. While an insomniac lauds late-night studying as "the beauty of online education," other women accustomed to more regular hours report that the third shift of education cuts into their already-scarce hours of leisure or sleep time. (Kramarae 2001, 33)

This is a very important point about the seeming advantage of online education for women. While it is true that women are able to get more education, which is a good thing, it is not changing the frameworks of their lives, which already assign them two "shifts" to men's one.

Despite the problems women have with access to online education, most that were able to participate in online courses were positive about the experience. Contrary to worries about the "digital divide," studies in developed countries such as Australia also find "no significant difference between female and male students with respect to being able to use the online learning environment confidently and effectively" (Coldwell, Goold, Craig, and Mustard 2007, 9). The same is true in the U.K. The Open University has been offering online and face-to-face versions of courses and collecting data on them for many years. Price studied the data from online and face-to-face versions of a course on technology (already biasing the result in some ways) and found that "The results show that they are confident independent learners who are academically engaged and may outperform their male counterparts online. Women do not have reduced computer and Internet access compared with men, nor are they disinclined to enroll on online courses" (Price 2006, 357-8).

Thus, women generally seem to see online education as an opportunity and to have the competence with computer technology to complete courses with good result. However, many of the problems that women with family and work responsibilities face in furthering their educations in face-to-face classes are not alleviated by online courses. Aside from this, there does not seem to be a gendered "digital divide"—the divide has more to do with education level and income. Those who already have both have access to online education. For those who do not, access is difficult.

3. Does online education have the potential to promote women's equality around the world?

One of the United Nation's Millennium Goals is to empower women through education (Empowering women through education 2004). The need for education in the "underdeveloped" world is enormous. In Asia alone, as many as 560 million adults are illiterate, the majority of them women. How, if at all, does online education help to meet that goal? And what are the barriers, if it does not?

The potential of online education for bringing about women's equality around the world is not great. The main obstacles are access to technology, the Internet, and electricity. Also, much of the education needed in the underdeveloped world is primary education, and online courses are not the best for this.

Just out of curiosity, I looked for some statistics of the availability of Internet access around the world (Table 1). They are striking.

So, many of the areas of the world where women are undereducated simply do not have access to the Internet. Furthermore, those organizations funding projects designed to provide such access always have profit in mind, and so tend to fund projects that will create consumers for their products. As Marlyn Tadros says in an article on Arab women and the Internet, "the internet continues to be an elitist tool whose access, cost, and skills make it prohibitive to many in the Global South, Arab women included" (Tadros 2005). Gulati agrees, arguing that computer technology has not contributed to equality, but has maintained existing hierarchies. For example, in India, access to computer courses designed for software engineers are highly competitive, sometimes expensive, and aimed at upper-middle class professionals. This merely maintains the status quo (Gulati 2008).

Besides sheer cost and the lack of infrastructure, there are other problems of access for people in underdeveloped nations. The dominant language of the Internet is English, and software in languages with non-Roman characters, such as Arabic, is very expensive. This means that the courseware available in developing countries often is in English. For example, the study of Greek women mentioned earlier said that most of the material in the program offered online was in English, because there is not much available in Greek.

There is some suggestion in the literature that women who are not allowed to attend mixed-sex classes might benefit from online education. An article on gender and distance education in Pakistan makes just this claim about distance education using paper materials (Hussein, Adeeb, Safdar, and Rahmanai 2008). Women in Pakistan are much less educated than men. There is some evidence that parents prefer their girls to be educated, but that they do not want them to leave home. Distance education, generally, has been useful to these girls, especially in areas of the country where women are "culturally restricted."

However, cultural gender differences make their way into online courses. Al-Harthi examined the experiences of students from the Arab Gulf taking online courses originating in the West.

Table 1. INTERNET USAGE STATISTICS

The Internet Big Picture: World Internet Users and Population Stats

World Regions	Population (2008 Est.)	Internet Users Dec. 31, 2000	Internet Users Latest Data	Penetration (% Population)	Users Growth 2000-2008	Users % of Table
Africa	975,330,899	4,514,400	54,171,500	5.6%	1,100.0%	3.4%
Asia	3,780,819,792	114,304,000	650,361,843	17.2%	469.0%	41.3%
Europe	803,903,540	105,096,093	390,141,073	48.5%	271.2%	24.8%
Middle East	196,767,614	3,284,800	45,861,346	23.3%	1,296.2%	2.9%
North America	337,572,949	108,096,800	246,822,936	73.1%	128.3%	15.7%
Latin America / Caribbean	581,249,892	18,068,919	166,360,735	28.6%	820.7%	10.6%
Oceania / Australia	34,384,384	7,620,480	**20,593,751**	59.9%	170.2%	1.3%
WORLD TOTAL	6,710,029,070	360,985,492	**1,574,313,184**	**23.5%**	336.1%	100.0%

NOTES: (1) Internet Usage and World Population Statistics are for December 31, 2008. (2) Detailed regional usage information on each world region is available at www.internetworldstats.com. (3) Demographic (Population) numbers are based on data from the U.S. Census Bureau. (4) Internet usage information comes from data published by Nielsen Online, by the International Telecommunications Union, by GfK, local Regulators, and other reliable sources. (5) For definitions, disclaimer, and navigation help, please refer to the Site Surfing Guide (www.internetworldstats.com/surfing.htm). (6) Information in this site may be cited, giving the due credit to www.internetworldstats.com. Copyright © 2001 - 2009, Miniwatts Marketing Group. All rights reserved worldwide.

She quotes a woman from the Gulf Arab States, who was in an online course with a man who knew her family:

> I've had a guy in one of my classes, and he's like my husband's friend. And whenever he enters the chat, I log off. …I don't feel comfortable especially a person, who knows us. Knows my husband. I just; I don't feel comfortable. Although I'm there to learn and there to participate, I'm not suppose to do that (laughs). It's very obvious when you log off. It shows, and when you log in, it shows like the guy saw me log off when he enters the chat, so, and the second time when I entered again he was gone, you know like it's like he felt that I logged off because of him. I don't feel comfortable. I just don't feel comfortable when a guy that knows us is there just I don't know why. …I'll be very nervous talking because I feel somebody is watching me, and probably making assumptions. … I'm trying to avoid you know, but I'm I'm not supposed; I'm not supposed to log off a chat because you know because I'm there. (Al-Harthi 2005, 8).

So, even though online education allows women who are restricted from mixed-group interaction to further their educations, the very cultural assumptions that restrict them emerge in online courses.

In her book on global eLearning, Alison Carr-Chellman discusses the concern that online education imposes Western ways of thinking and acting and ignores indigenous knowledge.

> First, inherent within what is often perceived as a value-neutral tool (the computer technology necessary for online learning) are a number of culturally biased amplifications which reinforce "cultural patterns of thinking that have their roots in the Industrial Revolution." (Reference omitted) (Carr-Chellman 2005, 8)

Secondly,

> …to fulfill the dreams of efficiency that make online education appealing to legislators and university administrations, customization and cultural sensitivity cannot be given adequate attention. Making a single course that is available around the world for anyone interested in it is efficient, but culturally and contextually bankrupt. …How can American professors, instructional designers, and Web educators realistically be expected to anticipate the cultural needs and contextual sensitivities necessary to create a course deliverable worldwide? (Carr-Chellman 2005, 9)

In reading articles on distance or online education and developing countries, I have often gotten a creepy feeling. I think that one of the elements of this feeling is that distance education/online education is seen as a quick and dirty means to development. All that matters is delivery—reception is not discussed. Remember our Greek adult learners.

Here is the introduction to an article that contains the creepiness:

> Advancement in distance education and electronic learning (DEEL) technologies and their widespread adoption are expected to result in significant socioeconomic changes across the world. In developing countries, these technologies have potential to reduce the gap between rich and poor, provide greater access to higher education for women, children and other socially disadvantaged groups, reduce population growth by raising the opportunity cost of bearing and raising children, and increase overall economic productivity. As such, strategies to promote DEEL opportunities should become part of socioeconomic development portfolios for local, regional, and national governments. (Alavalapati and Bannister 2007)

Another example: In a presentation on distance education for teacher education in Guyana, Nigeria, and Uganda, the author writes:

> Education in itself is regarded as a tool for development. In open learning there is the added value of the mechanisms and networks created to enable delivery, which also provide an artery right to the heart of the community. This socially constructed circulatory system can be tapped into and used to deliver other learning that might not be traditional distance education. (Binns 2002)

Add to this that many distance education projects are funded by international groups such as the World Bank. Online education is a product, and there are people making money from it. We have to watch out for this and make sure that neither dollars nor technology is driving educational opportunity.

Conclusion

In this survey of the topic of gender and online education, I have addressed three questions. First, what differences are there, if any, in the experiences of male and female students in online courses? I would say that there are differences, but that those differences do not seem to be inhibiting women's access to or success in online courses. There is some evidence that the online format can be particularly congenial to women, when properly constructed and monitored. Further, people who are reluctant to speak out in class are more willing to participate in semi-anonymous online discussion, and this category includes more women than men.

Secondly, online education seems particularly useful to women who have families and/or professional responsibilities, since it allows them to do their coursework when they can. However, this ease is somewhat misleading, since it does not really provide equal access with men who usually have only professional responsibilities in addition to coursework. It is reminiscent of the quotation about distance learning from the turn of the century: it allows women to "change their lives without altering or impairing the role society has sanctioned for them." In other words, the availability of online education may make traditional lives easier and more permanent. However, if women would not otherwise be able to educate themselves, online education is positive for women in the end. So, I would agree with the authors of a recent review of the literature on online education and gender, who stated that "research findings do not support the conception that women are disadvantaged" (Gunn, French, McLeod, McSporran, and Conol 2002, 33).

Will online education help to bring equality for women around the world? The cost of the infrastructure needed for Internet-based education, from electricity to connectivity, is out of reach for many. Those who can afford it are those who already have access to education. One group of women, denied education because they are not allowed into the spaces where it is provided, may benefit from online education. However, here, too, we see technology being used to preserve the status quo.

Some have commented on the anonymity that online courses can provide, but this seems to be in direct contradiction to the importance of self-disclosure that is often part of feminist pedagogy.[8] As Carr-Chellman says in her book on global e-learning,

> It is central to the construction of a viable democracy that diversity is recognized and celebrated rather than concealed. It is not inherent in online education. To advance either of these agendas, but rather the media lends itself more to concealment than to revelation. (2005, 4)

However, in most online courses, at least in the format with which I am most familiar, real names are used in postings so that many things about individual students might be (wrongly) assumed in accordance with stereotypes. But what happens when diversity is present, but ignored? How are those assumptions questioned? And, who benefits from this sort of situation?

Endnotes

1. See, for example, Hornsby and Maki 2008; Carpenter 2007; Brooke 2006; Boltuc 2005; Panza, Potthast, and Cathey 2006; Kemerling 1998; Carusi 2006; Heath 1998.
2. For other examples, see Gefen, Geri, and Paravastu 2007; Guiller and Durndell 2006; Sussman and Tyson 2000.
3. Hamann, Pollock, and Wilson (2003) found that "as gender mix approaches parity, females and males alike write longer messages, moderate their use of independent postings, and increase their reliance on direct or indirect interaction," all considered positive characteristics tending toward the "feminine" mode of communication. This is one of the few studies outside the education or technology fields. An "independent posting" is one that does not make reference to any other posting, and interactive postings are ones that do.
4. Some argue that online discussion can provide particularly safe places for female students. See Maher and Hoon 2008, 202: "If the physical space of the conventional classroom on campus is already shaped by the politics of gendered differences, will a pedagogical cyber domain change the ways gender shapes articulation and silence in the classroom? The answer we arrived at the end of our pedagogical journey is clearly 'yes'."
5. Monteith, Kirsteen. 2002. Gendered learning and learning about gender online: A content analysis of online discussion. University of Stirling. Last accessed September 30, 2009. http://www.odeluce.stir.ac.uk/docs/Gendered%20Learning.pdf
6. Cited in Bergman 2001, cited in Larreamendy-Joerns and Leinhardt 2006, 573-4.
7. See Vryonides and Zembylas citing Davies, P., Osborne, M., Williams, J. For Me or Not for Me? - That is the Question: A Study of Mature Students' Decision Making and Higher Education (Norwich: DfES, 2002).
8. Bray 2006, 169: "I believe that online teachers and students can create a more egalitarian, empowered and, therefore, productive educational relationship if they disclose and discuss their social diversity."

Women Don't Blog

H. E. Baber
University of San Diego

In our profession and other academic disciplines blogging has become the new hall talk. Philosophers operate their own blogs, post as members of group blogs, and enter into the discussion of philosophical and professional issues by commenting on posts. Women in the profession, however, are only half as likely to blog as their male colleagues. Women, I suggest, are reluctant to post because the risks of blogging for women are greater than they are for men and because they are less likely to benefit from assuming risk. If I am correct, gender dynamics in our profession induce women, as rational choosers, to play it safe when it comes to the decision whether to blog and, arguably, in making a range of other far more significant professional decisions.

Different Carrots, Different Sticks

There are a variety of reasons why academics should blog: Hugh McGuire, on his blog, lists nine, of which the following enjoy pride of place:

1. You need to improve your writing…blogging will help you get practice. 2. Some of your ideas are dumb. The sooner you get called out on bad ideas, the better…if you write a discipline-specific blog, then your colleagues around the world will read it…That means that when you have a dumb idea, you should hear about it quickly...when you have an incomplete idea, and some others chip in with suggestions, you'll get a better-formed idea.[1]

Most philosophers don't blog, and that should hardly be surprising. Many of us would rather not practice our writing in public or display dumb ideas to "colleagues around the world." There is, however, compensation for assuming the risk of putting one's unpolished drafts on display:

4. Blogging expands your readership…5. Blogging protects and promotes your ideas…6. Blogging is reputation.

Some philosophers are indeed willing to risk getting called out on dumb ideas in the interests of self-promotion and, more importantly, to engage in the public discussion that facilitates research. They blog: as members of group blogs, as commentators, and as proprietors of their own online enterprises. Within this group, however, women are significantly underrepresented. While there is no clear data on the percentage of philosophers who are women—estimates range between 17 and 30 percent[2]—the percentage of women in the profession who blog is significantly lower. Of the individual philosophy blogs maintained by philosophy faculty and students listed at David Chalmers' site, fewer than 10 percent of their owners are women.[3] Women are also disproportionately underrepresented amongst members of academic group blogs and less likely to comment at either group or individual blogs.

The underrepresentation of women on academic blogs is not restricted to philosophy and the cause has been the subject of widespread speculation. While I suspect there are a variety of factors at work, I shall suggest that the primary reason is that women are less able to afford risk than their male counterparts. As Donna Coker, commenting at Prawfsblog notes:

> [B]logging equals huge exposure… Lots of us law prof types are very risk averse in terms of what we publish. But women may have reason to be even more risk averse than men. Male law professors may believe (and they may be right) that mistakes they make in a blog post are not going to harm their credibility, while female law professors believe (and again they may be right), that their mistakes will indicate their lack of seriousness or their lack of intelligence.[4]

This is not to say that women are, in some global sense, more risk-averse, i.e., that given the same (perceived) costs, benefits, and probabilities of achieving a desired outcome women are more likely to play it safe. The claim is rather that women, with good reason, assess the same action a male would do as riskier: less likely to yield a good outcome, more likely to produce a bad outcome, and, if they don't achieve the desired result, likely to bring about a much worse outcome than they would for males who are on the face of it similarly situated.

Women have reason to worry because implicit bias which leads even people of good will to undervalue the work of women and members of other disadvantaged groups has been well-documented in over four decades of empirical research including, most recently, the administration of the Implicit Association Test (IAT) to large groups of subjects.[5] Members of disadvantaged groups are assessed less favorably than members of privileged groups. Women as well as men rate the same essays and professional resumes less favorably when women's names are attached. Blacks as well as whites exhibit the same implicit bias in favor of white males.

The results of this research come as no surprise to most women or minorities who have always known that they have to work harder, produce superior results, and, above all, be more careful than their white male counterparts in order to get comparable assessments of social acceptability and professional competence. White men can dress casually, behave boorishly, and talk like good ole boys without untoward consequences; black men know that they have to maintain a higher level of respectability to avoid being tagged as members of a criminal underclass, tailed by security guards, shunned, or hassled. Women know that they cannot afford to make mistakes. Men can afford to let their ideas, smart and dumb, fly without adverse consequences; women know that their dumb ideas are more likely to be taken as symptoms of incompetence or lack of seriousness, and that their smart ideas are less likely to be recognized and rewarded.

As a consequence, there are bigger sticks and fewer carrots for women in philosophy and other non-female-identified professions than there are for their male counterparts. Academic women know that they cannot afford to adopt McGuire's cavalier attitude about exposing dumb, or even incomplete, ideas to public scrutiny and, indeed, worry that blogging in and of itself may undermine their credibility. Writing anonymously on PrawfsBlog, an untenured legal scholar writes:

> Dan also raised issues that I am particularly sensitive to as a woman of color—blogging makes you incredibly visible (to the rest of the academy and to your own faculty). Certainly, it would be great to bring my work to the attention of those at other institutions, but the risk with blogging is that the blogging itself would be visible to my own faculty. I would be deeply worried (and perhaps I am overly paranoid) that blogging would be seen as "wasting time" I could be devoting to my scholarship. Further, the quick posts and responses typical in the blogosphere seem like they could come back to bite you. Suffice to say, I would not want blogging to come up as a centerpiece of my tenure review. Anyway, those are my thoughts. The fact that I didn't post [giving my name] suggests how paranoid I am about blogging, but there you are.[6]

If, as I have suggested, implicit bias figures significantly in professional evaluation, she has reason to worry. But even if it doesn't, the perception of implicit bias affects women's professional behavior: women are careful because they believe that implicit bias is at work. Whether rightly or wrongly, women believe that exposing dumb or incomplete ideas is very risky.

Women recognize also that their chances of having good ideas rewarded are relatively slim and so, arguably, blogging in order to protect or promote ideas or to build reputation has less appeal for women than for men who blog for glory. Again, the suggestion is not that women are inherently less ambitious, less interested in display or less interested in building "reputation" than their male counterparts but rather that they assume, again I believe with good reason, that their efforts are far less likely to be recognized and could, indeed, backfire.

Turning again to the illuminating discussion on PrawfBlog, Orin Kerr comments:

> I would flip the question and ask, what is it about men that attracts them to blogging? I would think the reason is that men love to hear themselves talk. They think they have something important to say; writing a

> lot becomes a way of showing off one's importance. Women usually have this baggage less often, so they're less likely to waste their precious few hours of free time in front of a computer blogging.[7]

Men, Kerr claims, like blowing their own horns. Women, he suggests, are less likely to carry the egoistic "baggage" that prompts male display and so are less likely to waste time blogging to show off.

This may or may not be true: it is moot whether, either by nature or nurture, women are less likely than men to believe that what they have to say is important or to have a taste for display. What I suggest is that whether or not women have a taste for display they are less likely to be rewarded for display and so will be less motivated to show off—particularly when there are significant risks. Women, because they believe, with good reason, that they will be judged more harshly than their male colleagues for careless quick responses, off the cuff remarks, and work that is not fully thought out or polished—the typical content of blog posts and comments—are cautious. And in a professional culture that values quick responses, boldness, and bravado, caution translates as mediocrity.

Women, indeed, are expected to be mediocre and rewarded for mediocrity: for being careful, competent, hardworking, tidy, dutiful, diligent, and solid—not brilliant or flashy. They are rarely rewarded for the bravado and swagger that are generally admired in bright young men, and may indeed be regarded as arrogant or "difficult to work with" if they show off or exhibit what is taken to be an inappropriately high degree of self-confidence. As Miranda McGowan notes:

> For many sex stereotypes…it's a short trip from description to prescription. For example, one common descriptive stereotype is that women are less competitive and more communal than men. It has a prescriptive side, too: people believe that "women should be communal and men should be agentic," that is, independent, individualistic, and competitive… Prescriptive stereotypes are enforced the same way as other social norms: violators suffer social sanctions, and both cross-typed men and women suffer. Women who are ambitious, self-promoting and competitive are perceived as unlikable and lacking social skills.[8]

Bloggers are expected to be both spontaneous and bold, to throw out unpolished work for discussion and critique. Given the nature of the medium it is of course possible for careful writers to fake it: it is possible, in principle, to contribute only posts and comments that are fully thought out, carefully articulated, and thoroughly polished. But this would be to defeat the whole purpose of posting on academic blogs—the quick and dirty discussion and critique of incomplete, unpolished works in progress—and would likely be a waste of valuable time and energy. If your work is polished and complete you may as well submit it to a journal—a much less risky procedure since submissions are blind reviewed by one or two referees and will not be exposed to the scrutiny of "colleagues around the world" under your name. Parting out your paper as blog posts or comments for quick, off the cuff responses by other participants is likely to be a waste of your time since, in revising and polishing, you have probably considered and responded to the quick and dirty objections. Comments by journal referees are more likely to be interesting and illuminating. On the other hand, if you post genuine work in progress, which is unfinished and unpolished, you risk serious embarrassment. This is a dilemma for all academics, not only in deciding whether to blog but when it comes to exposing works in progress to public scrutiny in any medium. For women, who can ill-afford to assume risk, the dilemma is starker.

In short, blogging is risky but the risks of blogging are greater for women than they are for men and the benefits are harder to come by. Consequently, women, acting as rational choosers, weighing the professional benefits, costs, and risks of blogging are less likely to blog than their male counterparts. This is an economic explanation in the broad sense of women's reluctance to blog and it is what might be called an "immediate-circumstance" explanation since it purports to explain a gender difference by reference to differences in the immediate circumstances of men and women. On this account, women in the aggregate behave differently from men, at least when it comes to blogging, because they are differently situated.

It remains to be seen how this explanation compares to competing explanations. I suggest that it is the most plausible explanation for most, even if not all, of the difference in the participation of men and women in academic blogs (1) because immediate-circumstance explanations for la difference are to be preferred to either nature or nurture explanations, (2) because it fits the data better than alternative explanations, and (3) because it explains other discrepancies in the behavior of men and women in the profession, including choice of specialties and research projects.

Why we should prefer immediate-circumstance explanations for *la difference*

In the aggregate, men and women play different roles in the family and in the labor force, behave differently, and make different choices. There are three different kinds of explanations available for these differences: biological/genetic explanations, "socialization" explanations, and immediate-circumstance explanations. It is likely that all of these factors figure in explaining gender differences. However, when looking for an explanation of differences in the ways in which humans behave we should, arguably, look first for immediate-circumstance explanations, then for explanations that appeal to socialization or early training, and last to biological explanations.

People respond to incentives

The purpose of this lexical ordering is not to "prove" in the teeth of empirical evidence that there are no biological-based psychological gender differences—there likely are statistical differences—or differences that are a consequence of early socialization, but to arrive at the best possible explanation of behavior. Humans are more similar to one another psychologically than they are to non-humans. Moreover, for all our interest in individual and cultural differences, human motivation and behavior is fairly consistent across culture and gender, and the assumption that individuals operate as rational choosers responding to incentives provides a decent, though imperfect, method for explaining and predicting behavior.

Such explanations are imperfect because homo economicus is an idealization. People are not ideally rational and do not operate with complete information. Behavioral economists challenge the neo-classical account and, incorporating empirical data from psychology and other social sciences, arguably, produces better explanations and more accurate predictions for human behavior. The fundamental assumption of such accounts, however, is the same: people respond to incentives and when explaining their behavior we should look first, and look hard, at the incentives to which they respond, the costs, benefits, and risks they consider in making decisions under uncertainty, before looking for "deeper" nature or nurture explanations.

This suggests that in explaining statistical differences in the way men and women behave, we should look for differences

in their immediate circumstances before reaching for socio-biological explanations or accounts that appeal to early socialization. Male philosophers are much more likely to post on academic blogs than women in the profession and, as Kerr suggests, this reflects a stronger preference for self-advertising. However, on the current account, before concluding that this is a consequence of "baggage" most men but relatively few women carry, presumably as a consequence of early socialization, we should look first for differences in incentives, differences in costs and risks men and women incur when they blow their own horns, and the benefits they might reasonably expect for engaging in this behavior.

Increasingly gendered differences in behavior are found to be a consequence of circumstance.

Citing an extensive body of research on gender differences, Miranda McGowan notes that "men and women increasingly exhibit and identify themselves as possessing the same set of personality traits":

> The Bem Sex-Role Inventory is a list of different personality traits, such as loyal, warm, assertive, analytical, happy, tactful, and jealous, which are categorized as expressive (feminine) traits, instrumental (masculine) traits, or neutral traits. A meta-analysis of college students' scores on the Bem Sex-Role Inventory shows that men and women's scores on instrumental or masculine traits have been steadily converging.[9]

Increasingly, men and women are similarly situated; increasingly, men's and women's scores converge, suggesting strongly that some psychological gender differences, which are commonly thought to be a consequence of early socialization or genetic hardwiring, are likely a response to differences in immediate circumstances.

Moreover, there is ample supporting data showing that in a range of cases where male and female behavior in the aggregate diverges, when data for men and women who are similarly situated are compared, the differences disappear. Consider, for example, the behavior of individuals in the labor forces. In the aggregate, women and blacks exhibit higher rates of absenteeism and quit behavior than white males. But looking for cultural, or biological, differences to explain this phenomenon would be seriously misleading since it turns out that when the figures are disaggregated, when women and blacks are compared to white men who do the same jobs the difference disappears. As it turns out and as we might expect, individuals who work at poorly paid, relatively unskilled, dead-end jobs are more likely take days off and to quit than those who work at better jobs with higher pay and more chance of advancement. And, as it turns out, women and blacks as a group are more likely to work at poorly paid, relatively unskilled, dead-end jobs than white men.[10]

Surveying the literature on gender differences, including the extensive body of work McGowan cites in her extremely useful paper, it seems clear that the search for immediate-circumstance explanations is fruitful. In addition, the reflexive appeal to early socialization or socio-biological explanations can lead us to overlook important differences in the immediate circumstances of men and women. And that leads us to the last and, from the moral point of view, most compelling reason to look first and look hard for differences in immediate circumstances when it comes to explaining differences in the behavior of men and women in our profession. If we assume that these differences are a consequence of biological differences or differences in early socialization we are likely to overlook differences in circumstance that are the result of implicit bias and skew the male-female playing field.

Pragmatic reasons to look for immediate-circumstance explanations first

Blogging is a matter of choice and, within our profession, women are only half as likely to men to blog. But when it comes to considerations of fairness, choice is not the end of the story. We make our choices in response to incentives and constraints. If the incentives and constraints men and women face are different that is unfair. And if those differences are a consequence of the way in which men and women are viewed and treated by colleagues it is a difference that we in the profession need to address in the interest of fairness.

We are supposed to treat likes alike. But if judgments of likeness and unlikeness depend on observations of behavior that is itself determined by the way in which people are treated the principle slips toward vicious circularity. If we treat people similarly, they are likely to behave similarly licensing further similar treatment; if we treat people differently, they will likely respond by behaving differently and so we shall infer that different treatment is warranted.

The discrepancy in the percentages of men and women in the profession who blog is striking, and exceedingly difficult to explain by reference to the small statistical differences in cognitive abilities and aggression that show up in psychological studies. Moreover, it seems likely that statistical psychological differences between men and women within the profession, selected and self-selected for similar aptitudes, interests, and (high) levels of aggression, should be smaller than male-female differences in the general population. In these circumstances, such a discrepancy is strong evidence for differences in the incentives and constraints that men and women in the profession face.

If such differences exist that is significant. Blogging is not a matter of great professional importance. Individuals' choice of specialty areas and decisions they make about the allocation of time to research, teaching, and other professional activities are. Men and women make different choices here, too. Within the profession, women are less likely than men to work in metaphysics and other central areas and more likely to favor teaching over research. A consideration of the reasons why so few women in the profession choose to blog, a relatively trivial matter, may illuminate the factors that induce women to make other professional choices that are far from trivial.

Professional choices

Within philosophy, and some other academic disciplines including law, women are less likely to blog than their male counterparts. I have proposed an explanation. Blogging, I suggested, is risky and the risks for women in academia are greater than the risks for their male counterparts. Moreover, I argued, for women the potential benefits are harder to come by. Women within the profession, as rational choosers, have stronger incentives than their male counterparts to play it safe and, as a consequence, are less likely to participate in academic blogs. More importantly, as I shall suggest, the structure of incentives which discourage women from blogging influence women's choice of specialties, decisions concerning their allocation of time between teaching and research and other professional decisions.

First I will consider my explanation of why women are less likely to blog than their male counterparts vis-à-vis competing explanations. Then I shall suggest that the same factors that discourage women in philosophy from blogging influence other, more significant, professional choices.

Blogging

One explanation for the dearth of women who participate in academic blogs is easy to dismiss. Women do not avoid blogging because of technical difficulties. Posting and commenting on blogs is easy and takes no more technical sophistication than sending email. Most academic business is now conducted on the Internet: journals and conferences prefer, or require, submissions by email or through websites. Academic women, like academic men, of necessity use the internet. As Margaret Crouch notes in her discussion of online education, reports of a male-female "digital divide, with males having more access to computers and greater competence, were overstated and currently does not seem to be a factor in the United States or other developed countries.[11]

A more plausible explanation for some of the discrepancy in male and female participation on blogs concerns the relatively low percentage of women who are members of group blogs. On most group blogs, posting privileges are by invitation only: to become a member of most group blogs you have to be tapped and women are far less likely to be tapped than men. Old boy networks and informal procedures notoriously favor white males over women and minorities, so, given the informal procedures for recruiting members for group blogs, it comes as no surprise that far fewer women are members.

This does not, however, explain why women are less likely to comment on blogs or why they run fewer professional and quasi-professional blogs themselves. Women are not tapped to be members of group blogs as often as men, but this does not explain all or even most of the difference between male and female participation.

Currently, one of the most popular explanations for differences in the choices men and women make appeals to women's "second shift," their responsibilities for housework and, most particularly, child care. Women do indeed work a second shift and that makes it more difficult for women in elite professions to get on the fast track and stay on. If it takes a 60+ hour per week work commitment to make partner in a glitzy law firm, then married women with children, who are de facto saddled with the major responsibility for housework and childcare, will be at a disadvantage relative to men and unmarried, childless women, and at an even greater disadvantage relative to men who have stay-at-home wives to handle housework, child care, and most of the business of life for them.

Nevertheless, though the second shift matters it does not explain most occupational segregation or all of the difference in choices men and women in academic occupations make. As McGowan notes,

> sex differences in interest don't explain occupational segregation. Women's supposed preference for more flexible schedules or a cushier work environment also explain little of the segregation. For one, many "pink collar" jobs are not flexible or family friendly. Retail jobs have inconsistent schedules that make scheduling childcare difficult, nurses work long hours and cannot telecommute. Low status workers usually have less control over when and where they can do their work than high status workers do.[12]

Likewise, it seems unlikely that lack of time, as a consequence of domestic responsibilities, accounts for women's reluctance to post and comment on academic blogs. At any given time, most women in the profession do not have young children and women who are students and so are less likely to have children and domestic responsibilities are no more likely to blog than other women in the profession.[13] Moreover, we don't see the same discrepancy in male and female participation when it comes to other forms of academic engagement, for example, participation in traditional conferences or other time-consuming professional activities. Lack of time may play some role in explaining why women are less likely to blog but it seems unlikely that that is the whole story.

I have suggested that the reluctance of women to participate in academic blogs is a consequence of women's rational risk-aversion. In the current section I argued that competing explanations are less plausible. I suggest further that the reasonable assumption by women that they cannot afford to assume risk influences their choice of specialties and other important professional decisions.

Other professional choices: a conjecture

In an ideal world, disinterested scholars would pursue research projects because they regarded them as important or, at the very least, interesting. In the actual world, few of us can afford to pursue disinterested scholarship. We select specialty areas with an eye to getting publications and write papers to get on APA programs, in order to add entries to our vitae. This is, as most of us will agree, a miserable business. The tail is wagging the dog. The purpose of publication is to make scholarly work that is interesting and important widely available in order to advance knowledge. But because, in order to get and keep jobs, we need to document professional activity we pursue research projects to get publications and are pressed to select topics not because they are either important or interesting but because they are likely to yield vita entries.

If I am correct, women are under more pressure than men to adopt this cynical policy because they are under more pressure to produce vita entries in order to be taken seriously and because they have no viable fallback positions, insofar as the prospects for women with generic humanities degrees outside of academia are dismal.[14] This poses a further question. Do women in philosophy tend to choose different specialty areas from men because they have different interests, because they are steered in different directions, or because they cannot afford to assume risk?

Sally Haslanger notes that "blatant discrimination [against women] has not disappeared…I know many women who have interests and talents in M&E who have been encouraged to do ethics or history of philosophy" and urges women to remember that they "have choices" and "don't have to put up with mistreatment."[15] I am not so sure about this. Trivially we do, of course, "have choices" but I am not so sure that most women have seriously viable options. We can choose to drop out and get secretarial jobs or, if we have the time and money, to start over again, get second, more salable BAs, MBAs, or law degrees. I seriously doubt, however, that most women in philosophy have, in any meaningful sense, choices.

This poses a serious question about women's choice of specialties. Women are underrepresented in metaphysics, epistemology, or other central areas in analytic philosophy. Does this reflect women's interests, or is it, as Haslanger suggests, a result of the tendency to steer women into other areas, or is it, as I suspect, a pragmatic decision to work in areas that are safer and less competitive, where it is easier to accumulate vita entries?

This is left as an open question—an exercise for the reader.

Fairness

If I am correct, women in philosophy make different decisions from their male counterparts because they cannot afford to assume risk. These are free choices insofar as they are rationally considered and voluntary. But it does not follow that treatment of women in the profession is just for if, as I have suggested,

women in the profession make the decisions they do because they are responding to different incentives and constraints than their mail colleagues, that is unfair.

In the literature on affirmative action it is customary to distinguish between "equality of opportunity" and "equality of result." We should certainly, so the story goes, work to achieve equality of opportunity for women and members of other disadvantaged groups. But if, given the opportunity, it turns out that there are proportionately fewer women or minorities than white males who are willing and able to occupy a range of positions, then there is no point in forcing unwilling or incompetent individuals to occupy these positions in the interest of achieving equality of result.

This is, however, a false dichotomy that arises from contrasting result with opportunity, which is not a matter of degree. The equality that matters for fairness in such cases is equality in costs, benefits, and risks. Women have the opportunity to blog or to pursue research in central areas of analytic philosophy. But, as I have argued, the costs and risks of pursuing these options are greater than they are for their male colleagues, and that is unfair. While there is nothing inherently desirable about equality of result, inequality of result is a symptom of inequitable differences in the incentives and constraints under which members of diverse groups operate.

Women in philosophy are far less likely to post on academic blogs than their male counterparts. There is nothing inherently wrong with this. It is, however, as I have suggested, symptomatic of differences in the incentives and constraints which men and women in the profession face. And that, arguably, is something which, in the interest of fairness, demands further consideration.

Endnotes

1. http://hughmcguire.net/2008/10/26/why-academics-should-blog/
2. http://en.wikipedia.org/wiki/Women_in_philosophy
3. http://consc.net/weblogs.html
4. http://prawfsblawg.blogs.com/prawfsblawg/2008/07/is-blogging-jus.html
5. http://www.projectimplicit.net/
6. http://prawfsblawg.blogs.com/prawfsblawg/2008/07/is-blogging-jus.html
7. http://prawfsblawg.blogs.com/prawfsblawg/2008/07/is-blogging-jus.html
8. Miranda McGowan. "Engendered Differences." Presented at University of San Diego Law School colloquium, May 1, 2009, pp. 38-9. In support, McGowan cites Alice H. Eagly, et al., "Transformational, Transactional, and Laissez-Faire Leadership Styles: A Meta-Analysis Comparing Women and Men" Psych. Bull. 129 (2003): 569, and Laurie A. Rudman & Kimberly Fairchild, "Reactions to Counterstereotypic Behavior: The Role of Backlash in Cultural Stereotype Maintenance," J. Personality & Soc. Psych. 87 (2004): 157, 163-64.
9. Miranda McGowan. "Engendered Differences." Presented at University of San Diego Law School Colloquium, May 1, 2009.
10. Francine D. Blau and Marianne A. Ferber. The Economics of Women, Men and Work (Prentice Hall, 2005).
11. Margaret Crouch. "Gender and Online Education" in this publication.
12. McGowan. "Engendered Differences," p. 23
13. See Chalmers' list of student bloggers at http://consc.net/weblogs.html
14. Vide, e.g., Robert Cherry. Who Gets the Good Jobs? Rutgers University Press, 2001. The male-female wage gap for non-graduates than it is for college graduates and sex segregation is higher for jobs that do not require a college degree. Women can avoid the worst effects of discrimination and sex segregation by getting special training and credentials. Without special training, credentials and qualifications women are more likely to be restricted to underpaid, dead-end, pink collar jobs. Men with "worthless" humanities degrees who drop out of academia had a reasonable chance of getting minimally decent white collar work; for comparably qualified women, without additional training and credentials, the options are strictly secretarial.
15. Sally Haslanger. "Canging the Ideology and Culture of Philosophy: Not by Reason (Alone)." Hypatia (2008).

Book Review

Social Networking Sites and the Surveillance Society. A Critical Case Study of the Usage of studiVZ, Facebook, and MySpace by Students in Salzburg in the Context of Electronic Surveillance

Christian Fuchs (Salzburg/Vienna: Research Group UTI, 2009). ISBN 978-3-200-01428-2.

Sandoval Marisol and Thomas Allmer
University of Salzburg

1. Introduction

"My data are sold for advertising," "connects people from all over the world and you find old and new friends," "Big Brother is watching you," "networking of students, exchange between like-minded people," "spying by employers," "entertainment and amusement" "international coming together," "the transparent human," "collaboration," "the surveillance society."

These are just some of the statements given by students asked about the advantages and disadvantages of integrated social networking sites (ISNS). The statements illustrate that students conceive social networking sites as contradictory: On the one hand, students see ISNS as possibilities for maintaining existing friendships; establishing new contacts; renewing old contacts; communicating, establishing, or maintaining international contacts; sharing photos and other media; and having fun. On the other hand, students stress risks of ISNS like political, economic, and personal surveillance; the possibility of employers to access profiles; advertising; unwanted contacts; stalking; harassment; and becoming a potential crime victim. Hence, "communication and surveillance are antagonistic counterparts of the usage of commercial social networking platforms" (96).

The ascertainment that social networking sites contain contradictory potentials is just one important result of a study published by Christian Fuchs, associate professor at the ICT&S Center of the University of Salzburg. His critical case study deals with the usage of studiVZ (studi = students, VZ = Verzeichnis = list; list of students), Facebook, and MySpace by students in Salzburg, Austria, in the context of electronic surveillance. In a first step the approaches of techno-pessimistic and techno-optimistic research about ISNS are criticized as forms of technological determinism and the author's own approach of critical research is developed. For doing so, Fuchs emphasize the social context of ISNS, namely, the political and economic interests of electronic surveillance in capitalist society and

concludes, "the only solution to privacy threats is to overcome new imperialism, surveillance society, and capitalism" (22). Subsequently, an empirical case study with almost 700 analyzed datasets is presented. It shows that students in Salzburg are rather less than knowledgeable but highly critical of the rise of a surveillance society. Students consider communication as the greatest opportunity and surveillance as the greatest risk of ISNS. Therefore, Fuchs deduces an antagonism between communication and surveillance in commercial social networking platforms and recommends, for instance, "to create non-commercial, non-profit social networking platforms on the Internet" (116).

The ability to describe social reality as contradictory and antagonistic shows Fuchs' association with critical theory as it has been founded by Karl Marx and has been advanced by representatives of the Frankfurt School. Thus, in order to discuss Fuchs' study we will first look at central characteristics of critical theory (section 2). In a next step we argue that Fuchs' study advances critical theory by applying it to contemporary social phenomena, and thus is an important contribution to critical theory in the information age (section 3). We conclude with some remarks on the overall value of Fuchs study for contemporary Internet research (section 4).

2. Elements of Critical Theory

The identification of antagonisms and contradictions, and the confrontation of existing social reality with its not-yet realized potentials are important elements of the critical theory of Karl Marx, Theodor W. Adorno, Max Horkheimer, and Herbert Marcuse.

In the "Preface to Contribution to the Critique of Political Economy" (Marx 1859, 7-11)[1] Marx stresses an antagonistic character of productive forces in capitalist society. On the one hand, relations of production control productive forces in the predominant conditions; on the other hand, "the productive forces developing within bourgeois society create also the material conditions for a solution of this antagonism" (Marx 1859, 9)[2]. Marx points at an oppressive character of existing social relations, but at the same time he identifies societal potentials for transcending the existing negativity. Hence, he is able to develop out of the existing actuality the true reality and new principles for the world out of the world's own principles, because "the material conditions for its solution are already present or at least in the course of formation" (Marx 1859, 9)[3]. Similarly, Adorno (1976, 68-70) states in The Positivist Dispute in German Sociology (1976, 68-70) that reality as it ought to be has to be confronted with the existing reality; consequently, criticism is necessary. Critical theory "must transform the concepts which it brings, as it were, from outside into those which the object has of itself, into what the object, to itself, seeks to be, and confront it with what it is. (...) In other words, theory is indisputably critical" (Adorno 1976, 69). It is a dialectic of essence and appearance. In Horkheimer's famous essay "Traditional and Critical Theory" (1982, 188-243), he stresses the necessity of thinking in social antagonisms; hence, he considers the contradiction of capital and labor and of productive forces and relations of production. Like Marx and Adorno, Horkheimer tries to show the real social possibilities, which result from advanced productive forces, and to develop an idea of the future: "Nonetheless the idea of a future society as a community of free men, which is possible through technical means already at hand, does have a content, and to it there must be fidelity amid all change" (Horkheimer 1982, 217). Also, Marcuse emphasizes in his essay "Philosophy and Critical Theory" that "current conditions and the analysis of their tendencies necessarily include future-oriented components" (1988, 145). So, he argues there are societal "potentialities that have emerged within the maturing historical situation" (1988, 158).

Why is critical theory able to describe society as contradictory and antagonistic, and why does it see both repressive and progressive developments at the same time? We argue that this results from several central characteristics of critical theory. The following list is certainly not exhaustive, yet the four elements described below allow explaining why critical theory looks at social phenomena as complex and contradictory, criticizes oppressive realities, and strives for emancipatory social change.

Dialectical Analysis:

Based on Hegel's dialectic, critical theory defines categories in relation to other things. Categories emerge in a dual way, cause, contradict, and negate each other; hence, it is a negation. Furthermore, raising quantity causes new qualities in dialectical categories at a certain critical point; hence, it is a turnover from quantity to quality. Finally, dialectical categories sublate each other. New qualities emerge, old ones are eliminated but are kept in a new form and on a higher level; hence, it is a negation of negation. Dialectical social criticism emphasize negations in society and supports a negation of negation for "a future society as a community of free men" (Horkheimer 1982, 217). It criticizes existing contradictory social conditions and asks for a cooperative society. On the ontological, epistemological, and praxeological level, dialectical philosophy considers social phenomena as complex, opposes one-dimensional thinking, and comprehends society as dynamic and changeable.

Society as Totality:

Critical theory has a certain term of "societal totality and its laws of movement" (Adorno 1976, 68). It detects a difference between essence and appearance and is able to identify reasons for social problems because it thinks in social totality. Thinking of society as a whole with its objective functions and developments is a crucial precondition for analyzing and criticizing society.

Humanistic Orientation:

Man is defined as a reasonable human being with happiness, self-determination, and liberty, and accordingly, as a Supreme Being. In capitalist society, man is alienated from himself where categories such as liberty are not realized. Critical theory is concerned about human beings and wants to liberate them because they are more than manipulateable subjects in the production process. Thus, critical theory has a humanistic and emancipatory character, "concern with human happiness, and the conviction that it can attained only through a transformation of the material conditions of existence" (Marcuse 1988, 135).

Historical Understanding:

Critical theory points out that the bourgeois mode of production is historically specific and changeable, not natural; thus, "the prehistory of human society accordingly closes with this social formation" (Marx 1859, 9).[4] Emancipation is not an idealistic idea, but a real materialistic possibility in contemporary society. Hence, critical theory emphasizes possibilities to transcend the existing negativity and develops transformative approaches.

Dialectical analysis allows critical theory to look at social phenomena as complex and contradictory; the consideration of society as totality allows to identify and to criticize power relations which shape certain social phenomena; because it is humanistic, critical theory wants to transcend the existing social reality and tries to foster emancipatory social change. Critical theory is able to conceive social reality as changeable because it looks at it as an historical result of specific human practices. These elements of critical theory can also be found in Fuchs' study on social networking sites.

3. Fuchs' Study as an Example for Critical Theory in the Information Age

Within contemporary Internet research Fuchs' study on social networking sites and surveillance is remarkable because it places the discussion on positive and negative effects of social software within a wider societal context. As he points out, most research on social networking sites is individualistic: it focuses on how individuals use ISNS either in a way that threatens them or in a way that empowers them. In contrast to such techno-optimistic and techno-pessimistic approaches, Fuchs' study is an example of critical Internet research, which updates critical theory and applies it to the study of contemporary social phenomena.

Dialectical Analysis:

As we have shown a central characteristic of critical theory of Karl Marx and the Frankfurt School is dialectical analysis. Dialectical analysis looks at social phenomena as complex and contradictory. It tries to identify contending pressures, contradictory forces, opportunities, and risks and shows to which extent these tendencies are realized and/or suppressed.

Dialectical analysis allows Fuchs to criticize techno-deterministic arguments of both techno-pessimistic and techno-optimistic approaches. For him both approaches are one-dimensional because they assume that technology has only one, either negative or positive, effect on society. In contrast, Fuchs looks at technology and society as "complex, dynamic systems" (13), which have "contradictory effects" (13). This means that technology contains the potential to be used in a repressive and/or emancipatory way. It can function as a means of exploitation and domination as well as a tool for strengthening the co-operative potentials of society. Societal effects of technologies thus depend on the societal context of their usage and can only be determined by analyzing underlying power-relations.

His dialectical approach allows him to avoid a one-sided view and to recognize that at the same time ISNS also contain positive potentials. He points out that ISNS support the maintenance of existing and the establishment of new friendships, community building, communicative exchange, and cooperation. However, in contemporary society the full realization of these emancipatory potentials is suppressed. The usage of ISNS is always accompanied by threats such as economic or state surveillance.

These contradictory potentials of ISNS are also reflected by the results of Fuchs' empirical study. One important result of the study is: "Although students are very well aware of the surveillance threat, they are willing to take this risk because they consider communicative opportunities as very important. That they expose themselves to this risk is caused by a lack of alternative platforms that have a strongly reduced surveillance risk and operate on a non-profit and non-commercial basis" (99). This shows that when using ISNS students are confronted with the contradiction that using these platforms at the same time brings advantages and poses threats.

The dialectical orientation of his approach also allows Fuchs to imagine an alternative that transcends the existing negativity. The transcendence is not located outside societal possibilities, and can be realized by transformative human practices. Transcendence in this approach is linked to immanent, material conditions of social reality. This allows defining a transcendent vision, which is not an idealistic utopia but a real social possibility.

By citing examples such as Wikipedia, Fuchs identifies tendencies that point beyond the existing social reality. The full realization of these potentials requires social transformations: "One needs to change society for finding solutions to problems. There are no technological fixes to societal problems" (14). The aim of a critical Internet theory is not only the establishment of the Internet as a public good and as a space for free and self-determined access to, exchange of, and co-operative production of information commons, but includes the transformation of society as totality.

Society as Totality:

Fuchs stresses that techno-pessimistic as well as techno-optimistic approaches on ISNS assume that societal risks and opportunities are inherent qualities of technologies. Fuchs disagrees with this assumption and points out that every discussion of risks and opportunities of ISNS has to consider the societal context.

This marks another important element of critical theory, which is a characteristic of Fuchs' study: to place the analysis of certain phenomena within the totality of society. He stresses the importance to "frame research issues by the macro context of the development dynamics of society as a whole" (21). The consideration of society as totality allows Fuchs to determine which of the contradictory potentials of technology is prevailing today, and how to foster the realization of emancipatory potentials.

Fuchs shows that ISNS are run by commercial enterprises and that "ISNS are objects of capital accumulation" (22). He argues that the main threats in regard to ISNS do not result from wrong behavior of individual users, but from corporate interests. For him profit-interests of new media corporations, which own ISNS, create the danger of state surveillance and economic surveillance: "On the one hand new imperialism has produced a situation, in which war and terror potentially reinforce each other, and the West reacts by increasing surveillance. […] On the other hand, not only the state, but also corporations have an interest in gathering personal data in order to develop personalized advertising strategies that target individual tastes and related tastes by aggregating and assessing user data" (33).

As Fuchs emphasizes these repressive effects of ISNS on individuals and society are not characteristics of technology as such but result from their usage by capitalist corporations. This insight can already be found in Karl Marx's "Capital." In the context of the industrial revolution and the rise of machinery he pointed out: "The contradictions and antagonisms inseparable from the capitalist employment of machinery […] do not arise out of machinery, as such, but out of its capitalist employment!" (Marx 1867, 465)[5]

Humanistic Orientation:

For a critical theory, which is humanistic and wants to foster the emancipation of humans and society, the question how threats and repressive realities can be challenged and advantages and emancipatory potentials can be realized is important.

Fuchs does not limit himself to pointing out the repressive realities and suppressed potentials, but he further develops ideas for transformative strategies that aim at fostering human emancipation. According to Fuchs, it is important to increase awareness of the repressive character of a capitalist usage of ISNS, which brings about threats such as economic and state surveillance. In order to increase critical knowledge Fuchs recommends to strengthen critical public discourse on surveillance, to organize information campaigns that show how people are immediately affected by surveillance, to document privacy violations, and to create non-commercial, non-profit social networking platforms on the Internet.

Historical Understanding:

Fuchs is able to identify societal alternatives because he

recognizes that the way in which ISNS are used today, the purposes they serve, and the effects they have result from specific human practices in contemporary society. This means that surveillance does not result from natural qualities of the technology, but from historical conditions and human practices. Technology could also be used in another way.

4. Conclusion

Fuchs' approach is rooted in Marxian philosophy and critical theory. It is a materialistic, dialectical, and historical approach, which is humanistic and interested in human emancipation and in the transformation of society as totality. We argue that this is a very promising background for studying the Internet and for assessing societal advantages and risks because it allows us to confront technological determinism and to look at technological structures and its effects as products of human practices and of societal power-relations.

The value of Fuchs' study does not only stem from the profound collection of empirical data on student's usage of ISNS, but also from its critical, dialectical orientation. This approach allows grasping the Internet as complex, contradictory, and as subject to contending pressures. Fuchs' analysis shows that repressive potentials are prevailing today. However, his approach allows him to identify suppressed possibilities, the tendencies that point beyond the existing reality, and the starting points for a transformation of the Internet and of society.

Fuchs' study is an important contribution to critical theory in the information age. It provides promising insights for scholars as well as for students who want to avoid technological determinism and to look behind mere appearances at societal power relations that shape technology and its usage.

Endnotes

1. English translation from: http://www.marxists.org/archive/marx/works/1859/critique-pol-economy/preface.htm (May 2, 2009)
2. English translation from: http://www.marxists.org/archive/marx/works/1859/critique-pol-economy/preface.htm (May 2, 2009)
3. English translation from: http://www.marxists.org/archive/marx/works/1859/critique-pol-economy/preface.htm (May 2, 2009)
4. English translation from: http://www.marxists.org/archive/marx/works/1859/critique-pol-economy/preface.htm (May 2, 2009)
5. English translation from: http://www.marxists.org/archive/marx/works/1867-c1/ch15.htm (May 2, 2009)

References

Adorno W., Theodor. 1976. Sociology and empirical research. In The Positivist Dispute in German Sociology, edited by Theodor Adorno W., Hans Albert, Ralf Dahrendorf, Jürgen Habermas, Harald Pilot, Karl R. Popper. 68-86. London: Heinemann.

Horkheimer, Max. 1982. Critical Theory. Selected Essays. London and New York. Continuum International Publishing Group.

Marcuse, Herbert. 1988. Negations. Essays in Critical Theory. London: Free Association Books.

Marx, Karl. 1859. Vorwort zur Kritik der Politischen Ökonomie. MEW, vol. 13, 7-11. Berlin: Dietz.

Marx, Karl. 1867. Das Kapital: Band 1. MEW, vol. 23. Berlin: Dietz.

SYLLABUS DISCUSSION

Teaching AI and Philosophy at School?

Aaron Sloman
School of Computer Science, The University of Birmingham

http://www.cs.bham.ac.uk/~axs/

1. Introduction: We Need Something Different

This paper proposes a way of teaching computing, not as a branch of engineering, but as a way of learning to do philosophy, cognitive science, psychology, linguistics, and biology, among other things. It could be the core of a new kind of liberal education. But what I am proposing is not new and untried—what is proposed is close to the spirit and philosophy of teaching programming and AI to complete beginners, which some of us developed at Sussex University from the mid 1970s onwards. A revival of that approach might address a serious current malaise. The vision presented here overlaps with that in Jeannette Wing's (2006), but has a different emphasis.

In the UK there has been much discontent in recent years about the teaching of computing in schools, not least because many bright learners form the impression that the study of computing is simply a matter of learning to use tools that everyone needs to learn to use, but without intellectual content of a type that could make it a subject worthy of study at university level. A similar view of chemistry might be produced if chemistry were taught mainly by teaching cooking. That is what has happened in schools by switching from the early experiments in teaching children to design, test, debug, and describe computer programs to teaching them only how to use word processors, email systems, web browsers, and possibly databases, spread-sheets and other tools—like attempting to teach physics by teaching pupils to drive cars and buses. This switch completely defeated the vision I wrote about in 1978:

> Another book on how computers are going to change our lives? Yes, but this is more about computing than about computers, and it is more about how our thoughts may be changed than about how housework and factory chores will be taken over by a new breed of slaves.
>
> Thoughts can be changed in many ways. The invention of painting and drawing permitted new thoughts in the processes of creating and interpreting pictures. The invention of speaking and writing also permitted profound extensions of our abilities to think and communicate. Computing is a bit like the invention of paper (a new medium of expression) and the invention of writing (new symbolisms to be embedded in the medium) combined. But the writing is more important than the paper. And computing is more important than computers: programming languages, computational theories and concepts—these are what computing is about, not transistors, logic gates or flashing lights. Computers are pieces of machinery which permit the development of computing as pencil and paper permit the development of writing. In both cases the physical form of the medium used is not very important, provided that it can perform the required functions.

> Computing can change our ways of thinking about many things, mathematics, biology, engineering, administrative procedures, and many more. But my main concern is that it can change our thinking about ourselves: giving us new models, metaphors, and other thinking tools to aid our efforts to fathom the mysteries of the human mind and heart. The new discipline of Artificial Intelligence is the branch of computing most directly concerned with this revolution. By giving us new, deeper, insights into some of our inner processes, it changes our thinking about ourselves. It therefore changes some of our inner processes, and so changes what we are, like all social, technological and intellectual revolutions. (From the preface.)

In pursuit of that dream, for many years (starting in 1974, in a team led by the late Max Clowes [Sloman1984b]), I was involved with teaching AI to novice university students; initially only students in humanities and social science disciplines, though, later, also students majoring in computing or AI. (NB: In those days hardly any first year students had even used a typewriter, let alone a computer.)

The idea was not that AI had produced theories about how minds worked, but that it provided a new way of thinking about what form good theories might take—e.g., not theories about necessary and/or sufficient conditions for some mental state to exist, as favored by most philosophers, nor theories about correlations between stimuli and behavior, as favored by most psychologists, nor theories about the advantages that might have led to the evolution of particular competences and traits, as favored by some evolutionary theorists, but theories about what minds could do and how they might do it, sought by designers and a few philosophers, e.g., Immanuel Kant.

Students learnt, from first hand experience of simple examples, about how explanatory theories can be developed, tested, debugged, extended, and in some cases used to generate new empirical research leading to better theories. My assumption was then (as explained in chapter 2 of the 1978 book) and remains now, that the alleged distinction between philosophy as a purely conceptual study and science as empirical was mistaken: when done well, philosophy and science overlap substantially. A well-known example is Einstein's thinking leading up to the special theory of relativity, using ideas from Hume and Mach (Sloman 1978, Chap. 3; Norton 2005). Computers provided powerful new tools to extend that overlap.

The approach to teaching, inspired by that vision, was developed with colleagues in the early years at Sussex University, from the mid 1970s and continued after I moved to Birmingham in 1991. We did not teach AI primarily as a branch of engineering, but as a way of trying to understand human (or, more generally, animal) competences, though with potential applications to engineering. No claim was made (by us) that the problems were close to being solved, or that human-like robots would soon be available [Sloman1978, Sec. 9.13, Chap. 10].

For some learners, their first ever programming exercise used the "riverworld" library with pre-built commands, such as **putin(X)**, **takeout(X)**, **getin()**, **crossriver()**, **getout()**, to instruct the computer to get a farmer, fox, chicken, and grain across a river, using a boat that could contain only two things, while avoiding ever leaving the chicken with the grain, or the fox with the chicken.

In Figure 1, colons precede user commands and asterisks precede program output. It illustrates that, although we had no graphical terminals in those days, the contents of a simple world could be displayed either in the form of a changing list of propositions stored in the program (the "database") or pictorially, using a pseudo-graphical display. The example also illustrates the friendly form run-time error messages ("Mishap messages") could take.[1] After an error, the program does not abort: interaction can continue, including the option of re-setting the world.

In another exercise, by exploring different ways of writing conversational programs, and seeing their limitations, students could begin deep new learning about the structure of their own language. Unlike many others teaching AI, we did not simply provide a programming language and expect students to build upwards from its primitive constructs. Instead, we produced a variety of library packages that we had written (taking full advantage of the support for advanced AI programming in Pop-11), and then allowed the students first of all to play with and use the packages, then to extend them, and then later to devise their own alternatives.

Figure 1. An example interaction with the Pop-11 "Riverworld" program.

```
: start();
** Here is the state of the river-world:
** [chicken fox grain man ---\ \_ _/ _____________ /---]
: database = = >
** [[boat isat left]
   [chicken isat left]
   [fox isat left]
   [grain isat left]
   [man isat left]]
: putin(grain);
** Here is the state of the river-world:
** [chicken fox man ---\ \_ grain _/ _____________ /---]
: database = = >
** [[grain isat boat]
   [boat isat left]
   [chicken isat left]
   [fox isat left]
   [man isat left]]
: getin();
** Here is the state of the river-world:
** [chicken fox ---\ \_ man grain _/ _____________ /---]

;;; MISHAP - DISASTER
;;; INVOLVING:  fox has eaten chicken TOO BAD
;;; DOING :  river_mishap eat checkeat getin pop_setpop_
   compiler
** Here is the state of the river-world:
** [fox ---\ \_ man grain _/ _____________ /---]
```

A student could try to design a grammar and lexicon for railway station announcements, and test it using the sentence generator provided in a library. Seeing some unexpected sentences, like "Platform 5 departed from the London train at 4pm" might lead the student to ask: How can I enrich my language specification so as to rule out sentences like that?[2] Our ideas were partly based on experience with the Logo programming language[3] and turtle graphics[4] using Logo, but after investigating Logo closely we decided a richer and more flexible language was needed (to which we later added an enriched turtle graphics subsystem). If we had not had a powerful programming language usable by both teachers and students we could not have developed such varied,

mind-stretching, teaching examples so easily, as explained in Sloman1984a.

2. Some philosophical issues

We also hoped our students would come to understand the differences between programs that could be described as blindly doing exactly what they were originally programmed to do, and programs that were able to modify themselves in the light of their "experience" so that what they did satisfied their own "preferences" rather than the preferences of their programmers. Likewise, they could change their preferences instead of being stuck with those provided by the programmer. This could give students new arguments to use in philosophical debates about whether machines could have free will and whether they could understand what they were doing and why, as opposed to being mere syntax manipulators. (See also [Ch. 2] Franklin1995.)

Another philosophical topic that emerges from thinking about how programs work involves the idea of a running virtual machine, as opposed to a virtual machine that is a mathematical abstraction whose instances are running virtual machines, e.g., the Linux virtual machine. In principle, this enables thoughtful students to have far more sophisticated discussions of problems about emergence, supervenience, and the mind-brain relationship than is common in philosophy, since most philosophers completely ignore the profound developments in computer science and software engineering that allow many virtual machines to run on their computers. Compare [Sloman Chrisley2003, Pollock2008, Sloman2008, Sloman2009].

Other important questions with philosophical implications that can fruitfully be discussed in the context of playing with, designing, or extending simple AI programs include questions about different forms in which information can be represented (like the pictorial and propositional representations of the river world), and questions about the implications of different information-processing architectures combining different sorts of interacting subsystems. Immanuel Kant, among others, might have benefited immensely from such experience.

Alas, neither learning about systems with human-like competences, nor learning new ways to think about old philosophical problems went on either in most university computer science or philosophy courses, or in most schools teaching computing.

3. Is it possible to start again?

As more and more computer power became available in schools, it seems there was more and more pressure to use computers simply as tools, supporting all sorts of tasks except the one thing that could have been of greatest educational value (for at least a subset of children): namely, learning to design, implement, test, debug, analyze, and explain working systems of increasing sophistication. The educational uses of computers in schools seem, for most learners, to have left out a key topic that is central to what makes computers possible, namely, the study of interactions between structures and processes, especially structures and processes in information-processing systems, such as minds and operating systems. Paradoxically, the problem many employers are complaining most about, namely, the lack of new graduate employees with programming skills, is a direct consequence of misguided decisions to teach children to use computers for the tasks that most employers, politicians, parents, and teachers thought computers were needed for, namely, using packages, not developing them.

A few years ago, when there was much agitation about how to get students to consider the study of computing as a worthy activity, I looked at some proposals for (re-)introducing programming ideas into the school syllabus, and felt that whereas they were well intentioned and would work well for some students, they would not attract some high fliers interested in the humanities, psychology, philosophy, etc.

A typical Computer Science (CS) syllabus takes a "bottom up" approach in which an understanding of computing is based on understanding of some of the important "low level" features of computers, and showing progressively how more complex capabilities can be built up. Such courses are intended to attract and teach students who will develop abilities needed to create computing systems that meet important current and future practical needs. However, it is wrong to assume that that is the only way, or even the best way, to inspire such learners. In any case, there are some very bright students whose ambitions are of a different sort and who would not be attracted by such CS courses, but nevertheless have the potential to acquire and use a very deep understanding of many varieties of computation. These students might go into other disciplines that require a deep understanding of varieties of information processing mechanisms.

So I proposed an alternative syllabus offering a type of education that would be attractive to students who were interested in the study of philosophy, psychology, language, social science, biology, or mathematics. The proposal sketched four modules to be studied in the last two years of school, two modules in each year, which could be studied alongside work in other disciplines.[5] It adopted a more "top down" approach, by introducing such students to "higher level" ideas such as the notion that there are various kinds of information processing, some of which occur in nature, e.g., in all animals, including microbes and humans, whereas others occur only in man-made machines. This contrast leads to the challenge of using man-made machines to produce systems that exhibit capabilities that are characteristic of humans and other animals. Such a syllabus should introduce:

- The high level goals of AI and Computational Cognitive Science, and some of its history[6];
- Some of the techniques and programming languages, and different forms of representation, that have been developed in pursuit of those goals;
- A brief introduction to some of the practical achievements of AI (including its growing importance in computer games and other entertainments)[7];
- Various kinds of challenges to AI, such as philosophical and empirical arguments claiming to prove that the goals are unattainable, and attempts at rebuttals;
- Ethical arguments related to whether the goals should be pursued and what the social and ethical consequences would be if they were achieved.

Whereas a conventional CS syllabus might start its programming component with work on logical and arithmetical operations, and demonstrate what could be built out of those, the proposed AI syllabus would begin with programs using a high level AI language along with a variety of libraries illustrating AI techniques, including list-processing techniques using pattern-based list construction and analysis.[8] Students could play with those libraries, initially using them, then combining them, then possibly extending or modifying them, and later developing new systems of their own, including simple kinds of "natural language" processing, such as generating and later parsing sentences, and then making plans, solving puzzles, and perhaps doing some learning, by keeping records, or modifying generalizations. Some schools might wish to introduce AI teaching based on programming physical or simulated robots, though that should be an option rather than a requirement. A toolkit supporting the development of simulated animals or

robots (e.g., "Braitenberg vehicles" [Braitenberg 1984]) with different internal architectures of varying complexity was added to Pop-11 in the decade after 1994.[9]

4. What is Artificial Intelligence (AI)?

AI is a (badly named) field of enquiry with two closely interrelated strands: science and engineering.

- The scientific strand of AI attempts to provide understanding of the requirements for, and mechanisms enabling, intelligence of various kinds in humans, other animals, and information processing machines and robots.
- The engineering strand of AI attempts to apply such knowledge in designing useful new kinds of machines and helping us to deal more effectively with natural intelligence, e.g., in education and therapy.

AI is inherently highly interdisciplinary because all kinds of intelligence, whether natural or artificial, are concerned with subject matters that are studied in other disciplines, and the explanatory models of natural intelligence have to take account of and be evaluated in the disciplines that study the natural forms.[10]

Like [Turing1950] I regard attempting to define "intelligence" as a waste of time. Instead, we can collect many different examples of competences displayed by humans or other animals, and examples of challenging biologically-inspired behaviors required in future machines, and we can investigate requirements for modelling or replicating them without needing to draw any definite line between those that are and those that are not intelligent. We may find it useful to subdivide the cases in terms of either their capabilities, or the mechanisms required, or the kinds of information they use, or their potential usefulness in various contexts. Those divisions will be much more interesting and useful than any binary division based on a pre-theoretical concept like "intelligence." Similar comments can be made about binary divisions between entities with and without consciousness [Sloman Chrisley2003], or with and without emotions [Sloman2001].

5. Why would students choose to study AI?

The collection of course descriptions below is aimed at students who are interested in finding out how important ideas associated with the development of computer-based systems are relevant to the broad study of naturally occurring information-processing systems, and to the development of new machines with human-like or animal-like capabilities.

Students taking such courses, sampling a variety of AI approaches and techniques, will not only start learning how to design, test, analyze, describe, and compare working computer models of diverse kinds, but will be better equipped than most to think about their broader significance in helping us understand such phenomena as human use of language, learning, development, visual and other forms of perception, problem solving, motive formation, and creativity [Boden 1990]. They may also learn new ways to think about evolutionary processes that produced such capabilities in humans and other animals. Students with an engineering bent can focus on some of the practical applications of these techniques, e.g., in medical diagnosis, in plant control systems, intelligent tutoring systems, computer games, and in new forms of entertainment. All students should be drawn into philosophical and ethical debates related to these ideas.

The course proposals below are not specifically aimed at students who wish to go on to higher education courses in computing or employment as computer developers or advanced computer users, though some of those students will certainly benefit from this unusual kind of computing education. It would also stretch the minds of students who wish to study other university subjects, such as psychology, biology, linguistics, philosophy, engineering, or management.

6. High level overview of a possible syllabus

Unlike the web site from which this proposal is derived,[11] this paper does not present a specific syllabus. The ideas presented here could be incorporated into very many different types of syllabus suited to learners of different ages with different backgrounds and career objectives. Merely in order to illustrate some of the possibilities I present a high level overview of a syllabus made up of four units, the first two of which might be taught during one year, and optionally followed by the second pair in the next year. It may be better to split some of the units into smaller, separately assessed components spread over a longer time. How much of a student's time the units would take would depend on other educational requirements: in the UK system it might be possible for students to spend between a quarter and about a third of their time on this work in their last two years. In other systems, requiring a wider spread of studies, the proportion of time would have to be reduced. Many modified versions of this outline might be offered to younger learners.

- **Unit 1:** Introduction to AI programming: building blocks

  Using simple, idealized models and games, students could learn to represent aspects of the world and rules of behavior in such a world. Factual and other information could be stored in list structures, using pattern matching where appropriate, with various kinds of procedures devised for constructing, comparing, analyzing, and interpreting different kinds of symbolic structures. Example demonstrations could include programs that analyze or generate sentences; hold simple conversations (initially Eliza-like, then more knowledge-based); draw, describe, and reason about pictures; explore simulated locations made of rooms, doors, and corridors; make simple plans; solve problems; and play games, where appropriate using pre-built libraries to provide some of the building blocks. Several of our students have enjoyed working on scenarios using simulated sheep and a sheepdog, with or without complex obstacles making the dog's task hard.[12] Others have investigated (simplified) models of motivation and emotion.[13]

  The programming techniques can make use of standard programming building blocks enhanced with AI mechanisms (e.g., use of local and global variables, conditionals, loops, recursion, case-constructs, along with pattern matching and rule-based programming). Simple AI toolkits may be used to implement concurrency, e.g., in simple adventure games or simulated robots or animals. Trainable neural net mechanisms might be available as libraries. For some students a basic introduction to logic programming could be included, e.g., using Prolog to manage and interrogate a simple database.

  Assessment could take many different forms, including, for example, use of interactive computer-based tests of understanding of programming constructs used in the course, individual mini-projects, and group mini-projects. Students should learn how to describe and compare working systems and should be able to write an essay on limitations and possible ways of extending something they have developed, or read about.

- **Unit 2:** History, philosophy, ethics, and social implications of AI

  Many essay and discussion topics could be related to the techniques encountered in Unit 1, including, for example, similarities and differences between symbolic AI, connectionist AI, and evolutionary computation. Students could learn enough to write about these without, at this stage, having learnt how to build all the varieties of AI programs. They should be able to answer questions about the importance of representations, algorithms, and architectures in AI systems, and the problems of choosing between alternatives. They should be able to explain the role of a running virtual machine as a platform on which competences can be built, and say a little about requirements for virtual machines to exist.

  They should be able to explain and illustrate the recurring problem of combinatorial explosions (in time and/or space requirements) and the differences between major complexity classes. They should know about some of the ambitious and controversial things being attempted in AI (e.g., attempts to give machines emotions), and understand some of the conceptual problems in defining such goals and evaluating progress, as well as some of the ethical problems.

  Regarding philosophical and ethical issues, there is a very varied range of possibilities here, including reading and discussing philosophical and ethical books and papers, or notes prepared by a teacher, with questions about whether computers could, in principle, replicate animal or human behaviors; whether machines could have experiences, motives, emotions, or values; whether development of such systems should be allowed; what we can learn about human nature from these investigations; what the implications are for evolutionary theories (e.g., evolution of intelligence, or unselfishness), what the long-term social, economic implications are; the ethics and practicalities of military applications of AI; and so on. Some students may wish to learn about possibilities and uses of realistic models of social interaction or socio-economic systems. Metaphysical problems to be discussed include the status of virtual machines, their components, and their causal relationships [Sloman2008, Sloman2009].

  There are many different forms of assessment possible, including essays, presentations, participation in debates, and writing short answers to some of the more technical questions.

- **Unit 3:** Advanced AI programming: designing integrated systems

  Depending on how much has been achieved in Unit 1, students could learn additional programming techniques, such as (depending on their interests or what their teachers can offer) creation of planners, reasoners, proof checkers, theorem provers, language understanding systems, conversation managers, neural nets, evolutionary computations, constraint nets, image analysis techniques, image interpretation, and tools for combining such components in an integrated working system, possibly a robot or simulated robot. Obviously, many of these will be possible only after several years of programming experience. However, it may be possible to learn by playing with and modifying or extending systems already developed by teachers or others. Everyone should have some experience of enabling disparate components to work together in an integrated architecture, possibly performing tasks concurrently, e.g., control of movement, perception, planning, communication and generation and evaluation of new motives. For younger, less experienced students this could mainly involve assembling pre-built units, and using existing integration toolkits (especially ones using object-oriented design with multiple inheritance, so that different functionalities can be composed).[14] Older students would include some components they have built themselves, as well as their own integration tools. Group projects would be very valuable in extending their communication and collaborative skills. There should be considerable emphasis on preliminary documentation of high level requirements as well as designs (e.g., using architecture diagrams), and on analytical or empirical comparisons of alternative solutions, as well as documentation of weaknesses and limitations of initial solutions to problems. In some cases, students can write about the philosophical, psychological, biological, social, or ethical implications of their work, or its possible extensions, though that would figure in Unit 4 if it is taught in parallel.

- **Unit 4:** AI Project

  The culmination of the learning in the various units could take the form of a group project, bringing together threads from the other units. Depending on local needs, resources, and timescales students might be able to devise their own project, or select from a list of possibilities provided by a teacher (with more or less specification detail provided in advance, and more or less of the required infrastructure provided in advance).

  It would be desirable to allow such projects to include use of physical robots, but that is not essential, and in many cases the use of carefully designed simulation environments can provide the most important kinds of learning—bypassing the important, but irrelevant problems connected with unreliable electrical or mechanical components. Other students, including those interested in intelligent fault analysis or design of robust systems, could use physical robots.

These hypothetical course units are merely illustrative of what is possible. Teachers should be allowed to use their independence and creativity so that they can tailor their teaching to their own expertise and interests, while meeting the needs of students and the broader community. Putting too much uniformity into a national syllabus can seriously deplete the pool of talents and ideas in the next generation as well as restricting opportunities for children who require unusual learning trajectories.

7. Prerequisites

Depending on the level and speed of presentation, students studying the earlier units will not require any prior knowledge of programming. A great deal of the work will involve typing text into a computer and reading textual output, and some students with visual or other disabilities may need special equipment or special help. Some will object that this is too difficult and learners should be given some of the recently developed graphical tools which allow children to develop working systems by assembling components using a computer

display and pointing device. Different tools are suited to different learners, but the view that everything should be made easy for learners is based on a seriously flawed understanding of the variety of cognitive transformations required in human learning. Only trivial things can be taught without generating confusion, so it is a mistake to try to avoid confusing students. However, teachers need to be able to help learners work through those confusions to deeper understanding—like learning to find your way around a town with an irregular street pattern caused in part by natural features like rivers and hills.

The proposed introductory units do not require specific mathematical skills, though understanding elementary arithmetic and logic will help. High logical and mathematical potential will be very useful, and the programming exercises should help to develop both, as well as providing opportunities to use such capabilities later on. For example, learning AI will inevitably involve learning some formal logic and set theory (both of which can be learnt as sub-tasks), and study of complexity issues can be used as a basis for teaching students about combinations and permutations. Requirements for programs with graphical interactions can be used to teach students about coordinate systems and some linear algebra. The most important prerequisite is a liking for solving problems with an intricate structure, such as crossword-puzzles, sudoku, or Rubik cubes, and a strong desire to understanding how complex things work.

8. Resources needed

It may be surprising to some people to learn how much can be achieved with relatively primitive and old-fashioned computing resources. When we started teaching at Sussex in 1974, we had to share a university mainframe computer that could not be used interactively, but from about 1975 we acquired our own PDP/11-40 computer with a small number of terminals (paper teletypes printing at ten characters per second!). As there was then no AI software available for that machine, Steve Hardy, appointed as a lecturer in AI, produced a reduced implementation of the Edinburgh AI language Pop-2,[15] which he called Pop-11.[16] That later grew into a multi-language system Poplog,[17] with incremental compilers for Pop-11, Prolog,[18] Common Lisp,[19] and ML,[20] first running on a VAX under VMS, then later ported to a variety of other machines, and sold commercially, mainly for AI teaching, research, and development (including software validation, plant control, expert systems, data-mining, and other applications). In 1998, ISL, the company then selling it for Sussex University, was bought by SPSS in order to take over the Clementine Data-Mining system, the most successful Poplog/Pop-11 product. After that Poplog became available as a free, open source system.[21] Despite its power and support for four major programming languages, the download package for version v15.63 on Linux requires under 17Mbytes, and the run-time system is very compact, enabling it to support a whole class of students sharing a single compute server.

This paper is not about Pop-11 or about Poplog, but the ideas proposed here were originally developed while using Pop-11 for teaching and research and have been used successfully, e.g., at Sussex University and at Birmingham University, some of it successfully using plain text visual displays, long before there were widely available graphical terminals. So these proposals are based on real experience of teaching courses somewhat like the ones being proposed. Although they were not taught to school children, as proposed here, some of them were taught to first-year arts and social studies students who had never previously used a computer. Moreover, at least one school teacher, Marcus Gray, demonstrated that the ideas could work in a British School [Gray1984]. Some of the tools and tutorials used at Sussex University, and elsewhere, are described on the poplog web site.[22]

9. Some practical challenges and possible (initial) solutions

There are many practical problems that will have to be addressed if a proposal like this is to be implemented on a large scale. Problems include: too few teachers, suitable programming tools not available in schools, lack of appropriate computing support staff, finding time in the syllabus (though I would argue that the educational value of this sort of activity could be more profound than several other things now taught in schools), funding, and time required to enable such a proposal to be developed and made available in a range of schools, with course materials suited to different ages and backgrounds. A potentially insuperable problem may turn out to be the declining interest in science and intellectual challenges in young learners, though courses of the kind proposed here might help to reverse that trend as well as showing young learners that computers are not merely (boring) tools that help you do "non-computational" things, like sending messages or downloading and playing music.[23]

Endnotes

1. For more details, see http://www.cs.bham.ac.uk/research/projects/poplog/teach/river.
2. See http://www.cs.bham.ac.uk/research/projects/poplog/teach/grammar for examples of the use of the "grammar" library with support for parsing and sentence generation using a student-supplied recursive context-free grammar.
3. http://el.media.mit.edu/Logo-foundation/logo/programming.htm.
4. http://en.wikipedia.org/wiki/Turtle_graphics.
5. Details are here: http://www.cs.bham.ac.uk/~axs/courses/alevel-ai.html.
6. Recently comprehensively surveyed in a two volume book by Margaret Boden (2006).
7. A vast amount of information about that, and other things, is available at this web site http://www.aaai.org/aitopics produced by the Association for the Advancement of Artificial Intelligence (AAAI).
8. For some examples, see http://www.cs.bham.ac.uk/research/projects/poplog/teach/matches.
9. http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html. A few examples of its use by university students are demonstrated in http://www.cs.bham.ac.uk/research/projects/poplog/figs/simagent. For use by younger learners simplified packages using the same tools could be provided.
10. Further information about the scope of AI is provided at http://www.cs.bham.ac.uk/~axs/courses/ai-overview.html.
11. http://www.cs.bham.ac.uk/~axs/courses/alevel-ai.html
12. See examples 3, 4, 5, and 8 here: http://www.cs.bham.ac.uk/research/projects/poplog/figs/simagent.
13. See examples 6, 9, 10, and 11 here: http://www.cs.bham.ac.uk/research/projects/poplog/figs/simagent.
14. Illustrated here: http://www.cs.bham.ac.uk/research/projects/poplog/teach/objectclass_example.
15. http://en.wikipedia.org/wiki/POP-2
16. http://en.wikipedia.org/wiki/POP-11
17. http://en.wikipedia.org/wiki/Poplog
18. http://en.wikipedia.org/wiki/Prolog
19. http://en.wikipedia.org/wiki/Common_Lisp
20. http://en.wikipedia.org/wiki/http://en.wikipedia.org/wiki/ML_(programming_language)
21. http://www.cs.bham.ac.uk/research/projects/poplog/freepoplog.html

22. http://www.cs.bham.ac.uk/research/projects/poplog/freepoplog.html#teaching
23. Further analysis of the problems and some partial solutions are discussed in these two online documents:

    http://www.cs.bham.ac.uk/~axs/courses/alevel-ai.html

    http://www.cs.bham.ac.uk/research/projects/cogaff/misc/compedu.html

References

[Boden 1990] Boden, M.A. 1990. The Creative Mind: Myths and Mechanisms. London: Weidenfeld & Nicolson.

[Boden2006] Boden, M.A. 2006. Mind as Machine: A History of Cognitive Science (Vols 1-2). Oxford: Oxford University Press.

[Braitenberg 1984] Braitenberg, V. 1984. Vehicles: Experiments in Synthetic Psychology. Cambridge, MA: The MIT Press.

[Franklin1995] Franklin, S. 1995. Artificial Minds. Cambridge, MA: Bradford Books, MIT Press.

[Gray1984] Gray, M. 1984. POP-11 for everyone. In New Horizons in Educational Computing, edited by M. Yazdani. 252-71. Chichester: Ellis Horwood Series In Artificial Intelligence.

[Norton2005] Norton, J. D. 2005. How Hume and Mach Helped Einstein Find Special Relativity. Synthesis and the Growth of Knowledge: Essays at the Intersection of History, Philosophy, Science, and Mathematics, edited by M. Dickson and M. Domski. Open Court (Forthcoming). (http://www.pitt.edu/ jdnorton/papers/HumeMach.pdf)

[Pollock2008] Pollock, J.L. 2008. What am I? Virtual machines and the mind/body problem. Philosophy and Phenomenological Research 76(2): 237-309. (http://philsci-archive.pitt.edu/archive/00003341)

[Sloman1978] Sloman, A. 1978. The Computer Revolution in Philosophy. Hassocks, Sussex: Harvester Press (and Humanities Press). (http://www.cs.bham.ac.uk/research/cogaff/crp)

[Sloman1984a] Sloman, A. 1984a. Beginners need powerful systems. In New Horizons in Educational Computing, edited by M. Yazdani. 220-34. Chichester: Ellis Horwood Series In Artificial Intelligence. (http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#45)

[Sloman1984b] Sloman, A. 1984b. Experiencing computation: a tribute to Max Clowes. In New Horizons in Educational Computing, edited by M. Yazdani. 207-19. Chichester: Ellis Horwood Series In Artificial Intelligence. (http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#71)

[Sloman2001] Sloman, A. 2001. Beyond shallow models of emotion. Cognitive Processing: International Quarterly of Cognitive Science 2(1): 177-98.

[Sloman2008] Sloman, A. 2008. Virtual machines in philosophy, engineering & biology [Extended abstract]. Proceedings Workshop on Philosophy & Engineering WPE-2008, edited by N. McCarthy and D. Goldberg. Royal Academy of Engineering, London. (http://www.cs.bham.ac.uk/research/projects/cogaff/08.html#803)

[Sloman2009] Sloman, A. 2009. What cognitive scientists need to know about virtual machines. Proceedings of the 31st Annual Conference of the Cognitive Science Society, edited by N.A. Taatgen and H. van Rijn. Austin, TX: Cognitive Science Society. (http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#901)

[Sloman Chrisley2003] Sloman, A. Chrisley, R. 2003. Virtual machines and consciousness. Journal of Consciousness Studies 10(4-5): 113-72.

[Turing1950] Turing, A. 1950. Computing machinery and intelligence. Mind 59: 433-60. (Reprinted in Computers and Thought, edited by E. A. Feigenbaum and J. Feldman. 11-35. McGraw-Hill, New York, 1963.)

[Wing2006] Wing, J.M. 2006. Computational Thinking. CACM 49(3): 33-35. (http://www.cs.cmu.edu/afs/cs/usr/wing/www/publications/Wing06.pdf)

CALL FOR PAPER WITH *ETHICS AND INFORMATION TECHNOLOGY* ON "THE CASE OF E-TRUST: A NEW ETHICAL CHALLENGE"

Trust in digital environments (e-trust) affects the activities of millions of individuals involving a wide range of social dynamics. This pervasive phenomenon raises new ethical problems, such as the occurrence of e-trust relationships between human and artificial agents and the emergence of trust in on-line contexts.

The ethical debate on e-trust has been characterised by the tension between two opposite positions. One considers e-trust as a different phenomenon from trust. It argues that trust requires embodied interactions characterised by emotional, cultural and physical aspects and hence that trust could not arise in digital contexts, where such kinds of interactions are impossible. The other position rejects the assumption of embodied interactions as a necessary condition for the occurrence of trust, and focuses on the analysis of the main characteristics and of the ethical features of e-trust.

The purpose of this special issue of Ethics and Information Technology, entitled "The Case for e-Trust: a New Ethical Challenge," is to address explicitly the issues concerning the ethical nature of e-trust.

Submitted papers are requested to explore issues concerning the following research questions:

1. What are the fundamental and distinctive aspects of e-trust?
2. Should e-trust be regarded as an occurrence of trust on-line or as an independent phenomenon in itself?
3. What are the ethical implications of e-trust?
4. To what extent artificial agents can be involved in an e-trust relationship?
5. What is the influence, if any, of the context on the emergence of e-trust?

Submissions will be double-blind refereed for relevance to the theme as well as academic rigor and originality. High quality articles not deemed to be sufficiently relevant to the special issue may be considered for publication in a subsequent non-themed issue of Ethics and Information Technology.

The editorial project is officially endorsed by the UNESCO Chair in Information and Computer Ethics.

Closing date for submissions: March 1st, 2010.

To submit your paper, please use the Springer online submission system, to be found at www.editorialmanager.com/etin

Please contact the special guest editors for more information,

Mariarosaria Taddeo, mariarosariataddeo@gmail.com

Luciano Floridi, luciano.floridi@philosophy.ox.ac.uk

Or the managing editor,

Noëmi Manders-Huits

N.L.J.L.Manders-Huits@tudelft.nl