NEWSLETTER | The American Philosophical Association

Philosophy and Computers

FALL 2013 VOLUME 13 | NUMBER 1

ISSN 2155-9708

APA NEWSLETTER ON

Philosophy and Computers

PETER BOLTUC, EDITOR

VOLUME 13 | NUMBER 1 | FALL 2013

FROM THE EDITOR

Peter Boltuc
UNIVERSITY OF ILLINOIS–SPRINGFIELD

We are lucky, and for more than one reason. First, we were able to secure an important article, one of the most serious defenses of the inconsistency theory of truth. It is so far the main paper that came out of John Barker's Princeton dissertation that became pretty famous already in the late 1990s. Barker's conclusion (closely related to classic arguments by Chihara and based primarily on the liar paradox) is that the nature of language and the notion of truth, based on the logic of language, is inconsistent. Sounds like Plato's later metaphysics in J. Findlay's interpretation, doesn't it? Then, at the last moment, Dan Kolak brought an important article by Jaakko Hintikka. While Dan introduces Hintikka's paper in his note from the chair, let me just add my impression that this is one of Hintikka's most important works ever since it highlights the potential for function logic. Hence, we have two featured articles in this issue. Just like John Pollock's posthumous article in theory of probability for AI (artificial intelligence; this newsletter, spring 2010), those are works in which philosophy lays the groundwork for advanced computer science.

Second, we have a brief but meaningful note from Tom Powers, the incoming chair. When I joined this committee ten years ago, it was led by Marvin Croy and a group of philosophers, mostly associated with the Computers and Philosophy (CAP) movement. Members were very committed to advocating for various uses of computers in philosophy, from AI to online education. All of us were be glad to meet in person at least twice a year. We had active programming, sometimes two sessions at the same APA convention. Then we would meet in the evening and talk philosophy at some pub until wee hours. And yes, the chair would attend the meetings even if his travel fund had been depleted. I have a strong feeling that under Tom's leadership those times may be coming back, and soon.

We are also lucky to have a number of great articles directly linked to philosophy and computers in this issue. Keith Miller and Dave Larson, in their paper that caused great discussion at several conferences, explore the gray area between humans and cyborgs. John Basl, in a paper written in the best tradition of analytical moral theory, explores various ethical aspects of creating machine consciousness.

It is important to maintain a bridge between philosophers and practitioners. We are pleased to include a thought-provoking paper by Christophe Menant, who discusses many philosophical issues in the context of AI. We are also glad to have two outstanding papers created when the authors were still graduate students; both were written for a seminar by Gordana Dodig-Crnkovic. Linda Sebek provides a hands-on evaluation of various features of assistive environments while Juan Durán discusses philosophical studies of computer simulation. I would like to encourage other educators in the broad, and necessarily somewhat nebulous, area of philosophy and computers to also highlight the best work of their students and younger colleagues.

FROM THE CHAIR

Dan Kolak
WILLIAM PATERSON UNIVERSITY

I am happy to report that we have, in this issue, a fantastic follow-up (of sorts—a more apt phrase might be "follow through") to Jaakko Hintikka's previous contribution, "Logic as a theory of computability" (*APA Newsletter on Philosophy and Computers*, volume 11, number 1). Although Jaakko says of his latest piece, "Function Logic and the Theory of Computability," that it is a work in progress, I am more inclined to call it a "progress in work."

Had my little book *On Hintikka* (2011) been written two decades earlier, it would have consisted mainly of accounts of his early work on logic—Hintikka's invention of distributive normal forms for the entire first-order logic, his co-discovery of the tree method, his contributions to the semantics of modal logics, inductive logic, and the theory of semantic formation. Instead, I had to devote most of the space to the then-recent past twenty years. To summarize his work in the dozen years since would take an entire new book. (That I am not alone in this assessment is evidenced by the Library of Living Philosophers bringing out a second Hintikka volume.) Indeed, when John Symons and I, in *Questions, Quantifiers and Quantum Physics: Essays on the Philosophy of Jaakko Hintikka* (2004), considered the importance of Hintikka's work, we said, half tongue in cheek, that its philosophical consequence is not the additive property of the sum of its parts, and used an analogy: "Hintikka's philosophical legacy will be something like the philosophical powerset of his publications and lines of research."

Being chair of the APA committee on philosophy and computers for the past three years has been a wonderful learning experience. Although it has become a truism that most interesting things happen at the borders, nowhere is this most clearly evident than at the intersection of philosophy and computers, where things that develop faster perhaps than at any other juncture tend to be consistently,

refreshingly, often surprisingly, and dangerously deep. Nowhere is this more evident than in this newsletter, which under the insightful and unflappable stewardship of Peter (Piotr) Boltuc has been functioning, often under duress, as a uniquely edifying supply ship of new insights and results. Peter deserves great credit and much thanks. By my lights he and this newsletter are a paradigm of the APA at its best. Thank you, Peter, and happy sailing!

FROM THE INCOMING CHAIR

Thomas M. Powers
UNIVERSITY OF DELAWARE

The official charge of the APA committee on philosophy and computers describes its role as collecting and disseminating information "on the use of computers in the profession, including their use in instruction, research, writing, and publication." In practice, the committee's activities are much broader than that, and reflect the evolution of philosophical interest in computation and computing machinery. While philosophy's most direct connection to computation may have been through logic, equally if not more profound are the ways in which computation has illuminated the nature of mind, intelligence, language, and information. With the prominent and growing role of computers in areas such as domestic security, warfare, communication, scientific research, medicine, politics, and civic life, philosophical interest in computers should have a healthy future. Much work remains to be done on computers and autonomy, responsibility, privacy, agency, community, and other topics.

As the incoming chair of the committee on philosophy and computers, I want to encourage philosophers to make use of the committee to explore these traditional and new philosophical topics. I also invite APA members to suggest new ways in which we as a profession can deepen our understanding of computers and the information technology revolution we are experiencing. Please consider contributing to the newsletter, attending committee panels at the divisional meetings, suggesting panel topics, or nominating yourself or others to become members of this committee.

ARTICLES

Truth and Inconsistent Concepts

John Barker
UNIVERSITY OF ILLINOIS–SPRINGFIELD

Are the semantic paradoxes best regarded as formal puzzles that can be safely delegated to mathematical logicians, or do they hold broader philosophical lessons? In this paper, I want to suggest a philosophical interpretation of the liar paradox which has, I believe, nontrivial philosophical consequences. Like most approaches to the liar, this one has deep roots, having been first suggested by Tarski (1935) and later refined by Chihara (1979).[1] I offered a further elaboration of the idea in *The Inconsistency Theory of Truth* (1999), and here I would like to develop these ideas a bit further.

The term "liar paradox" refers to the fact that the ordinary disquotational properties of truth—the properties that allow semantic ascent and descent—are formally inconsistent, at least on the most straightforward way of formally expressing those properties and given standard assumptions about the background logic. The best-known formulation of those disquotational properties is Tarski's convention (T):

(T) "*A*" is true if and only if *A*

We now consider a sentence such as

(1) Sentence (1) is not true.

As long as the schematic letter *A* in (T) has unlimited scope, we can derive the following instance:

(2) "Sentence (1) is not true" is true if and only if sentence (1) is not true.

Then, noting that the sentence quoted in (2) is none other than sentence (1) itself, we derive the consequence

(3) Sentence (1) is true if and only if sentence (1) is not true.

And this conclusion, (3), is classically inconsistent: it is an instance of $P \leftrightarrow \sim P$.

The liar paradox should concern all of us, because it represents a gap in our understanding of truth, and because truth is a central notion in philosophy, mathematical logic, and computer science. Tarski's (1935) work on truth is what finally put mathematical logic on a firm foundation and led to the amazing explosion of work in that field. Tarski's work in turn inspired Davidson (1967), whose influential work gives truth a central place in semantic theory. And computer science, of course, is based on mathematical logic; the theory of computability itself is essentially just the theory of truth for a certain fragment of the language of arithmetic.[2] (For more on the relation between logic and computability see Hintikka's (2011) contribution to this newsletter.) If truth plays such an important role in all three fields, then it behooves us to get to the bottom of the paradoxes.

There is now a truly vast body of literature on the liar, and the argument (1–3) above is far from the last word on the subject. Having said that, the liar paradox is remarkably resilient. Accounts of the liar can be divided into two camps: descriptive and revisionary. For a revisionary account, the goal is to produce a predicate with disquotational properties of some sort, which can serve the purposes that we expect a truth predicate to serve, while not necessarily being wholly faithful to our naïve truth concept. This approach has much to recommend it. But in this paper, I will focus on *descriptive* accounts. If the ordinary notion of truth needs to be replaced by a revised notion, I want to know what it is about the ordinary notion that forces us to replace it. If the ordinary notion is defective in some sense, I want to know what it means to say it is defective. And if, on the other hand, we can produce an account of truth that avoids contradiction and is wholly faithful to the ordinary concept, then there is no need to go revisionary.

Descriptive accounts, in turn, can be divided into the following categories, depending on what they hope to achieve.

- **Block the contradiction.** Descriptive accounts in this category proceed from the assumption that there is a subtle but diagnosable flaw in the reasoning that leads to contradictions such as (3). Indeed, it's not hard to convince oneself that there *must* be such a flaw: if an argument has a contradictory conclusion, there must be something wrong with its premises or its inferences.

- **Embrace the contradiction.** On this approach, there's nothing wrong with the reasoning leading up to the conclusion (3). That conclusion simply expresses the fact that the liar sentence (1) is both true and not true. This approach, known as dialetheism,[3] has never been the majority view, but lately it has received a surprising amount of attention.

- **Acknowledge the contradiction.** On this approach, convention (T) is part of the meaning of "true," and so the contradiction (3) is in some sense a consequence of the concept of truth. This differs from "embracing" the contradiction in that the contradiction (3), while viewed as a commitment of ordinary speakers, is not actually asserted. This will be the approach taken here.

Revisionary accounts also try to block the contradiction; and if the contradiction can be effectively blocked, then doing so is the preferred approach, I would think. But blocking the contradiction turns out to be hard, especially (I will argue) in the context of a descriptive account. In the next section, I will explain some of the reasons why this is the case. If blocking the contradiction is as hard as I think it is, we should at least entertain the alternatives, provided the alternatives are intelligible at all. In the remainder of this paper, I will try to explain what it means to acknowledge the contradiction, and why it makes sense to do so.

1. WHY THE LIAR IS HARD

Any account of the liar, whether descriptive or revisionary, has to operate within the following constraint:

Constraint 1. The truth predicate, as explained by the theory at hand, must have the expected disquotational properties.

And this by itself is not easy to achieve: we saw earlier that a natural formulation of the "expected disquotational properties" led directly to a contradiction. Having said that, there is some wiggle room when it comes to "expected disquotational properties," and we also have some leeway in our choice of background logic. In fact, there are theories of truth that have some claim to satisfying Constraint 1.

Let's consider a couple of examples: not the highest-tech examples, to be sure, but sufficient for our purposes. First, Tarski's original proposal was to simply restrict convention (T) so that the substituted sentence *A* is forbidden from containing the truth predicate. Then the substitution of sentence (1) for *A* is prohibited, and the contradictory conclusion (3) cannot be derived. But this restriction on (T) is quite severe, limiting what we can do with the resulting truth predicate even in a revisionary account. For a descriptive account, Tarski's restriction is simply a non-starter, since natural language clearly places no such limit on what can substitute for *A* in (T). (And it should be noted that Tarski himself viewed this approach as revisionary, not descriptive.)

Another approach to revising (T), which results in a less severe restriction, starts from the idea that not all sentences are true or false. In particular, some sentences represent truth value gaps, with the liar sentence (1) a very plausible candidate for such treatment. If gaps are admitted, then we can maintain an equivalence between the sentences *A* and "*A* is true" for all *A* in our language. In particular, when *A* is gappy, so is "*A* is true." The first mathematically rigorous treatment along these lines is due to Kripke (1975), who developed a family of formal languages containing their own gappy truth predicates, each obeying a suitable version of (T). Sentences like (1) can then be proved to be gappy in Kripke's system.

The main weakness of Kripke's approach is that the languages in question need to be developed in a richer metalanguage. Some of the key notions of the account, while expressible in the metalanguage, are not expressible in the object language. In particular, the notion of a gappy sentence, which is obviously crucial to the account, has no object language expression. The reason is simple and instructive. On the one hand, in Kripke's construction, there is an object language predicate *Tr*, and it can be shown that *Tr* is a truth predicate in the sense that (a) an object language sentence is true if and only if it belongs to *Tr*'s extension, and (b) an object language sentence is false if and only if it belongs to *Tr*'s anti-extension. (Predicates in Kripke's system have extensions and anti-extensions. A predicate *P* is true of those objects in its extension, false of those in its anti-extension, and neither true nor false of anything else.) Now suppose the object language had a gappiness predicate as well. That is, suppose there were a predicate *G* whose extension included all and only the gappy sentences. We could then construct a sentence that says "I am either not true or gappy"—i.e., a sentence *S* that is equivalent to $\sim Tr(\ulcorner S \urcorner) \vee G(\ulcorner S \urcorner)$. *S*, like any sentence, is either true, false or gappy. But if *S* is true, then both $\sim Tr(\ulcorner S \urcorner)$ and $G(\ulcorner S \urcorner)$ are not true, and thus neither is *S*. If *S* is false, then $\sim Tr(\ulcorner S \urcorner)$ is true, and thus so is *S*. And if *S* is gappy, then $G(\ulcorner S \urcorner)$ is true, and hence so is *S*. So *S* is neither true, false, nor gappy, which is impossible. This contradiction (in the metatheory) proves that no such predicate as *G* exists.

Kripke described this phenomenon as the "ghost of the Tarskian hierarchy" because despite his efforts to create a self-contained object language, he found it necessary to ascend to a richer metalanguage, just as Tarski did. The problem is also called the *strengthened liar problem* because the sentence *S* is a "strengthened" (i.e., harder to deal with) version of the liar sentence, and also as the *revenge problem*, since the moment we account for one manifestation of the liar problem, a new manifestation appears to take revenge on us. The key feature of the revenge problem is that in addressing the liar we develop a certain set of conceptual tools (in this case, the notion of a truth value gap). Those tools are then turned against us—i.e., they are used to construct a new liar sentence (in this case, *S*) which our original account is unable to handle.

Whatever we call it, the revenge problem shows that even though Kripke was able to construct an internally consistent way of satisfying truth's expected disquotational properties, he did so at the expense of placing a tacit restriction on the sorts of sentences that the resulting truth predicate applies to. Specifically, he constructed a truth predicate for a language in which the metalanguage notion of gappiness is inexpressible. The construction used to create the strengthened liar sentence *S* is rather general, and the *prima facie* lesson of the revenge problem is that an account of truth can't be given for the language in which the account is formulated.

If this is so—and so far it has been suggested but not proved—then it is moderately bad news for revisionary accounts and extremely bad news for descriptive accounts. From a revisionary perspective, the revenge problem simply means that in constructing a predicate with the desired disquotational properties, we will have to be content with a predicate that applies only to a certain fragment of the language we speak. Some sentences in our language may be assertible, and we may even be committed to asserting them, but we can't use our (revisionary) truth predicate to describe them as true: they simply fall outside that predicate's scope. This might be a limitation we can live with. But from a descriptive perspective, it is puzzling. The ordinary concept of truth applies, or at least it certainly seems to apply, to all sentences of our language, not just to some formally tractable fragment of our language. That is, descriptive accounts have to live with the following additional constraint.

Constraint 2. A descriptive account of truth must describe a truth predicate for an entire natural language, not just a fragment of a natural language.

So suppose we have an account of truth, and suppose it uses some notion, like gappiness, that doesn't occur in the sentences to which the truth predicate, as described by our theory, applies. In what language is this account stated? The natural obvious answer is that it is stated in a natural language (e.g., English). But then what we have produced is an account of truth for a proper fragment of English, not for all of English, in violation of Constraint 2.

For this reason, it has often been suggested that when we formulate an account of truth, we sometimes do so not in an ordinary language like English, but in a richer language, call it English+.[4] English+ is English supplemented with technical terms, like "gappy," that are simply not expressible in ordinary English. And the resulting account is a theory of true sentences of English, not of English+.[5] Such a move faces some challenges, however.

First of all, if one holds that English+ is needed to formulate a theory of truth for English, then it is hard to resist the thought that a still-further enhanced language, English++, could be used to formulate a theory of truth for English+. The process can clearly be iterated, leading to a sequence of ever-richer extensions of English, each providing the means to express a theory of truth for the next language down in the hierarchy. We can even say exactly how this works: English+ comes from English by adding a predicate meaning "gappy sentence of English"; English++ comes from English+ by adding a gappy-in-English+ predicate; and in general, for each language *L* in the hierarchy, the next language *L*+ is obtained from *L* by adding a predicate for the gappy sentences of *L*.

However, once we have all this on the table, a question very naturally arises: What language are we speaking when we describe the whole hierarchy of languages? Our description of the hierarchy included the fact that English+ has a predicate for gappiness in English but "gappy in English" is not expressible in English, so our account must not have been stated in English. Parallel reasoning shows that our account cannot have been stated in *any* language in the hierarchy. We must have been speaking some super-language English* that sits at the top of the entire hierarchy. And then we're right back where we started, since clearly we need a theory of truth for English* as well.

Maybe a better approach is to just drop talk of the hierarchy of languages, or at most to understand it as a form of Wittgensteinian gesturing rather than rigorous theorizing. But there is another problem. Let's just focus on the languages English and English+, where again English+ is the result of adding a predicate to English that means "gappy sentence of English." English+ is, again, the metalanguage in which we diagnose the liar paradox as it arises in English. This approach assumes that the truth predicate of English applies only to sentences of English: English has a predicate meaning "true sentence of English," but does not have a predicate meaning "true sentence of English+." If it did, then that predicate could be used to construct a gappiness predicate in English. Specifically, we could define "gappy sentence of English" in English as follows:

> *A* is a gappy sentence of English if and only if the sentence "*A* is gappy" is a true sentence of English+.

And since English does not have a gappy-in-English predicate—the entire approach depends on this—it doesn't have a true-in-English+ predicate either. More generally, if English had a true-in-English+ predicate, then English+ would be translatable into English, which is impossible if English+ is essentially richer than English. So any theory of truth that, by its own lights, can only be stated in an essentially richer extension English+ of English must also maintain that (ordinary) English lacks a truth predicate for this extended language.

All of this sounds fine until one realizes that the truth predicate of English (or of any other natural language, I would think) is not language-specific. The truth predicate of English purports to apply to propositions regardless of whether or not they are expressible in English. This should actually be obvious. Suppose we discovered an alien civilization, and suppose we had good reason to suspect that the language they speak is not fully translatable into English. Even if we assume this is the case, it does not follow that the non-translatable sentences are never used to say anything true. On the contrary, it would be reasonable to assume that some of the extra sentences are true. But then there are true sentences that can't be expressed in English. Or suppose there is an omniscient God. Then it follows that all of God's beliefs are true; but it surely does *not* follow that all of God's beliefs are expressible in English.

So the ordinary truth predicate applies, or purports to apply, to sentences of any language, and this fact forms another constraint on descriptive accounts:

Constraint 3. The truth predicate, as described by the account, must apply to sentences of arbitrary languages (or to arbitrary propositions).

But this constraint is incompatible with the richer-metalanguage approach. To see this, suppose "gappy sentence of English" really is expressible only in some richer language English+. This means that some people—some philosophers who specialize in the liar, for example—actually speak English+. Let Bob be such a speaker. That is, let "Bob" be a term of ordinary English that denotes one such speaker. ("Bob" could abbreviate a definite description, and there are plenty of those in ordinary English.) Then we can say, in ordinary English, for any phoneme or letter sequence *A*,

(4) The sentence "*A* is gappy" is true in Bob's idiolect.

If "true" behaves the way it intuitively seems to, as described in Constraint 3, then (4) is true in English if and only if *A* is gappy in English. So English has a gappiness predicate after all, which directly contradicts the account we have been considering.

For these reasons, I think an account of truth that requires a move to a richer metalanguage is unpromising as a descriptive account, however much value it might have as a revisionary account. So what are the prospects for a descriptive account that does not require a richer metalanguage? A complete answer would require a careful review of the myriad accounts in the literature, a monumental undertaking. But let me offer a few observations.

First, because the problem with expressing gappiness is a formal problem, it is relatively insensitive to how the gaps are interpreted. Because of this, numerous otherwise attractive proposals run into essentially the same revenge problem. Here are some examples.

Truth is a feature of propositions, and the liar sentence fails to express a proposition.

This is an attractive way of dealing with liar sentences, until one realizes that failing to express a proposition is just a way of being gappy, and that the usual problems with gappiness apply. The strengthened liar sentence, in this case, is

(5) Sentence (5) does not express a true proposition.

Does sentence (5) express a proposition? First, suppose not. Then a fortiori, (5) does not express a true proposition. In reaching this conclusion, we used the very words of (5): we wound up assertively uttering (5) itself. And in the same breath, we said that our very utterance failed to say anything. And our account *committed* us to all this. This seems to be an untenable situation, so maybe we should reconsider whether (5) expresses a proposition. But if (5) does express a proposition, then that proposition must be true, false, or gappy (if propositions can be gappy), any of which leads to trouble. Here's another example:

> There are two kinds of negation that occur in natural language: wide-scope and narrow-scope (or external and internal). In the liar sentence (1), the negation used is narrow-scope. When we step back and observe that (1) is not true, our "not" is wide-scope.

Well and good, but the natural and obvious response is to simply construct a liar sentence using wide-scope or external negation:

(6) Sentence (6) is not_{wide} true.

Then, in commenting that (6) is gappy and thus not true, we are assertively uttering the same words as (6) in the very same sense that was originally intended.

A perennially popular response is to regard truth ascriptions as ambiguous or otherwise context-sensitive and to diagnose the liar on that basis.[6] The intuition behind this response is as follows. We would like to say that (1) is gappy, and being gappy is a way of not being true. So we reach a conclusion that we express as follows:

(7) Sentence (1) is not true.

Formally, sentence (7) is the same as the liar sentence (1), and so in assertively uttering (7), we are labeling the words of our very utterance as not true. Intuitively, though, there seems to be an important difference between the utterances (1) and (7). In (7), we are stepping back and evaluating (1) in a way that we weren't doing with (1) itself. This has led some philosophers to suggest that (1) and (7) actually say different things.

The tools to formally express this idea go back to the Tarskian hierarchy of languages and, before that, the Russellian hierarchy of types. Using Burge's (1979) account as an example, suppose we explain differences like that between (1) and (7) in terms of differences in the content of "true" on different occasions. That is, suppose we treat "true" as indexical. Let's use numerical subscripts to mark the different extensions of "true": true_1, true_2, ... Then sentence (1), fully subscripted, is rendered as follows:

(1′) (1′) is not true_1.

On an account like Burge's, (1′) is indeed not true: i.e., it is not true_1. We express this in the same words as (1′):

(7′) (1′) is not true_1.

But in assertively uttering (7′), don't we commit ourselves to the truth of (7′)? Indeed we do, but not to the truth_1 of (7′). From (7′), what we are entitled to conclude is

(8) (7′) (and thus (1′)) is true_2.

And there is no conflict between (7′) and (8). Problem solved! A bit more formally, what we have done is modify the disquotational properties of truth somewhat. We have, for any given sentence *A* and index *i*,

(T_i^1) If "*A*" is true_i, then *A*

And we have a weak converse: for any *A*, there exists an index *i* such that

(T_i^2) If *A*, then "*A*" is true_i

This modified disquotational principle is perfectly consistent, and on the face of it, it leaves us with a perfectly serviceable disquotational device.

One question that can be raised about such a proposal is whether there is any evidence, aside from the paradoxes themselves, that the natural language word "true" really works this way. I do think this is a worry, but there is another, potentially more damaging problem. Consider the following sentence, sometimes called the "super-liar":

(S) Sentence (S) is not true_i for any *i*

Using (T_i^1), it is easily seen that (S) is not true_i for any *i*. That is, (S) is not true *at all*: there is no context in which it is correct to say that (S) is true. And yet our conclusion here—sentence (S) is not true_i for any *i*—is stated in the very words of (S), so there had better be some sense in which (S) is true. Thus, we have what seems to be a violation of (T_i^2).

The standard response is that (S) is simply ill-formed: it relies on binding the subscript *i* with a quantifier, which is not permitted. This response is correct as far as it goes, but it misses the fact that (S) *is* a well-formed sentence of the metalanguage in which the account is presented. Or at least, something with the same gist as (S) can be expressed in the metalanguage. After all, the account at issue makes explicit generalizations about the hierarchy of truth predicates, for example the claims (T_i^1) and (T_i^2). Such claims presuppose some mechanism for generalizing across indices, and once that mechanism is in place, we can use it to construct sentences like (S). Indeed, (S) and (T_i^1) are entirely parallel: each is (or can be written as) a schema with a schematic letter *i*, understood as holding for all indices *i*. If you can say (T_i^1) in the metalanguage, you can say (S) too.

But we plainly can't say (S) in the object language, so we're back to the problem of the essentially richer metalanguage. Notice also that the problem of (S) is a classic example of the revenge problem: the machinery of the account—in this case, the ability to generalize across indices—is used to construct a new liar sentence that the account can't handle.

In summary, we have found some substantial obstacles to a satisfactory descriptive account of truth, at least if that account is to satisfy the three constraints mentioned above; and those constraints are certainly well-motivated. What are we to make of this?

2. THE INCONSISTENCY THEORY

One possible response to these considerations is to simply reject one or more of Constraints 1-3. However, there are different things that it can mean to reject a constraint. It might be that at least one of the constraints is simply factually wrong: the natural language truth predicate doesn't work like that, even though it seems to. Alternatively, we could argue that while the constraints are in fact part of the notion of truth, there is no property that satisfies these constraints, and hence, no such property as truth. My proposal will be somewhat along the latter lines, but let's first consider the former proposal.

One could certainly reject one or more of the constraints of the last section as factually incorrect, but such a move seems to me to be very costly. Suppose, for example, that we reject Constraint 1, that truth has the expected disquotational properties. For example, suppose we maintain that in some special cases, assertively uttering a sentence does not carry with it a commitment to that sentence's truth. This would free us up to assert, for example, that

(9) (1) is not true

without worrying that this will commit us to the truth of (9) (and hence, of (1)): the above sentence may simply be an exception to the usual disquotational rule.

But one seldom finds such proposals in the literature, and I think the reason is clear: the disquotational principles seem to be part of the meaning of "true." One might even say they seem analytic. And this consideration seems to have a lot of pull, even with philosophers who don't believe in analyticity. Finding a sentence that turns out to be an exception to the disquotational rules would be like finding a father who is not a parent. The disquotational rules seem to me to be so much a part of our notion of truth that rejecting them would be tantamount to declaring that notion empty.

Likewise, one could question whether a descriptive theory needs to apply to the language it's stated in. That is, one could reject Constraints 2 and 3. But this would be tantamount to claiming that the ordinary notion of truth applies only to a proper fragment of the language we speak, or at least a proper fragment of a language we could (and some of us do) speak, and it seems clear that truth, in the ordinary sense, has no such limitation.

Yet another possibility is to simply accept the existence of truth value gluts: of sentences that are both true and not true. This at least has the virtue of simplicity. Convention (T) can be taken at face value and there's no need for complicated machinery or richer metalanguages. As for the costs of this approach, many would consider its commitment to true contradictions to be a cost in itself.

But suppose we could get the explanatory benefits of dialetheism without being saddled with true contradictions. That is, suppose there were a way to maintain that (T), or something like it, really is part of the concept of truth without actually claiming that liar sentences are both true and untrue. Such an account might be very attractive.

Along these lines, let's start with a thought experiment. Imagine a language where nothing serves as a device of disquotation. The speakers get together and decide to remedy the situation as follows. First, a string of symbols is chosen that does not currently have a meaning in the language. For definiteness, let's say the string in question is "true." Next, the following schema is posited, with the intent of imparting a meaning to this new word:

(T) "*A*" is true if and only if *A*.

It is understood that *A* should range over all declarative sentences of the language, or of any future extension of the language. And that's it: positing (T) is all our speakers do to impart any meaning or use to "true." The word "true" goes on to have a well-entrenched use in their language long before anyone realizes that contradictions can be derived from (T).

There are a number of observations we can make about this thought experiment. First, it is coherent: we can easily imagine a group of speakers doing exactly what I have described. We can certainly debate what *meaning*, if any, the word "true" has in their language, but it seems clear that a group of speakers could put forward (T) with the *intention* of giving a meaning to the new word "true."

Second, we can easily imagine that the positing of (T) leads to "true" having a well-defined *use* in the speakers' language. We simply have to imagine that "is true" is treated as a predicate and that the application of (T) as an inference rule becomes widespread. We might even imagine that once the use of "true" becomes well-entrenched, the explicit positing of (T) fades from memory—but that's getting a bit ahead of the story.

Third, in saying that the speakers establish a use for "true," we should understand "use" in a *normative* sense, as governing the *correct* use of "true," and not just as summarizing speakers' actual utterances or dispositions to make utterances. This is crucial if we want to say that (T) has a special status in the language and isn't just a pattern that the speakers' behavior happens to conform to. It is also the sort of thing we should say in general: the notion of use that is relevant to questions of meaning, I claim, is the normative sense. In any case, I think it's clear from the thought experiment that (T) is put forward as a norm and adopted as a norm by the speakers.

Fourth, I claim that the positing and subsequent uptake of (T) confers a meaning on "true," in *some* sense of "meaning." Here we have to be careful because the word "meaning" itself has several different meanings, and "true" (in this example) may not have a meaning in every sense. It's not obvious, for example, that "true" has a well-defined intension. What I mean is that "true" in the imagined case is not simply nonsense; it plays a well-defined role in the language.

Fifth, and finally, there is nothing in this thought experiment that forces us into dialetheism in any obvious way, even if we accept the foregoing observations. We've simply told a story about a language community adopting a certain convention involving a certain word; doing so shouldn't saddle us with any metaphysical view about things being both so and not so. To put it a bit differently: there's nothing contradictory in our thought experiment in any obvious way, so we can accept the scenario as possible without thereby becoming committed to true contradictions. Of course, the speakers themselves are, in some sense, committed to contradictions, specifically to the contradictory consequences of (T), but that's a separate matter. There's a big difference between contradicting yourself and observing that someone else has contradicted herself.

It should come as no surprise that I think the above thought experiment bears some resemblance to the actual case of the word "true" in English. However, there is an important difference between the two cases. Namely, no natural language ever got its truth predicate from an explicit positing of anything like (T). We shouldn't read too much into this difference, however. In the thought experiment, the initial stipulation of (T) plays an important role, but an even more important role is played by the speakers' incorporation of (T) into their language use. Eventually, the fact that (T) was stipulated could fade from memory, and any interesting feature of the word "true" would depend on its ongoing use. In which case the question arises: What interesting feature does "true" have in these speakers' language?

The best answer I know is that the speakers have a language-generated commitment to (T), which was initially established by the act of positing (T) and then sustained by the speakers' ongoing use of "true." I think this accurately describes the language of the thought experiment, and I suggest that (aside from the business about positing) it describes natural languages as well. In the case of natural language, (T) is not an explicit posit, but it is a convention of language, accepted tacitly like all such conventions.

So this is the inconsistency theory of truth as I propose it. In natural languages, there is a language-generated commitment to the schema (T) or something very much like it. Using (T), we can reason our way to a contradiction. This gives rise to the liar paradox, and it explains why the liar is so puzzling: we don't know how to block the reasoning that generates the contradiction because the reasoning is licensed by our language and our concepts themselves.

As evidence for the inconsistency theory, I would make the following points. First, the considerations of the previous section should make an inconsistency theory worth considering. Second, the inconsistency theory is simple: no elaborate gyrations are required to avoid paradox, either in our semantic theory or in the conceptual schemes we attribute to ordinary speakers. And third, the inconsistency theory does justice to the sheer intuitiveness of (T). My native speaker intuitions tell me that (T) is analytic, and the inconsistency theory supports this intuition. Indeed, if one were to accept the inconsistency theory, it would be very natural to *define* a sentence to be analytic in a given language if that language generates a commitment to that sentence.

The inconsistency theory shares these virtues with dialetheism, which is unsurprising given the similarity of the two views. But (as I will argue at greater length in the next section) the inconsistency doesn't actually have any contradictory consequences. For those philosophers (like me) who find true contradictions a bit hard to swallow, this should be an advantage.

3. REFINEMENTS, OBJECTIONS, AND RAMIFICATIONS

Is the inconsistency theory any different from dialetheism, though? We need to know, that is, whether the inconsistency theory implies that the liar is both true and not true, or, more generally, whether it implies both *P* and not *P* for any *P*. Equivalently, we need to know whether the inconsistency theory is an inconsistent theory.

One might argue that the present account makes logically inconsistent claims about obligations. On our account, we have a language-generated commitment to (T). This means that at least in some circumstances, we have an obligation to assert (T)'s instances, as well as the logical consequences of (T)'s instances. Thus, we have an obligation to assert that the liar sentence (1) is true, and we also have an obligation to assert that (1) is not true. Now if the logic of negation also generates a prohibition on asserting both *A* and not *A*—as I think it does—then we have a case of conflicting obligations. And, it can be objected, this latter claim is itself inconsistent.

What this objection gets right is that the inconsistency theory regards the language-generated commitment to (T) as a kind of obligation and not (or not just) as a kind of permission. It's not that we are licensed to infer *A* from "*A* is true" and vice versa, but need not make this inference if we don't feel like it: if we assert *A*, we are thereby committed to "*A* is true," and are therefore obligated to assert "*A* is true," at least in those circumstances where we need to express a stance on the matter at all. Moreover, the obligations in question are unconditional: they have no hidden escape clauses and can't be overridden like Ross-style *prima facie* obligations.

The only proviso attached to the commitment to (T) is that it is conditional upon speaking English, and specifically on using "true" with its standard meaning. We can always use "true" in a nonstandard way, or even refrain from using it altogether, working within a "true"-free fragment of English. The point of the present account is that *if* we choose to go on using "true" with its ordinary meaning, then we are thereby committed to (T).

So is it inconsistent to say that a given act is both obligatory and prohibited? For whatever reason, this matter seems to be controversial, but I think there are many cases where conflicting obligations of just this sort clearly do occur.

Case 1. A legislature can create a law mandating a given act *A*, or it can create a law prohibiting *A*. What if it (unknowingly) did both at once? Then the act *A* would be both obligatory and prohibited under the law.

Case 2. People can enter into contracts and thereby acquire obligations. People can also enter into contracts with multiple third parties. What if someone is obligated to do *A* under one contract, but prohibited from doing *A* under a different contract?

Case 3. Games are (typically) based on rules, and a poorly crafted set of rules can make inconsistent demands on the players. As a simple example, imagine a variation on chess—call it chess*—with the following additional rule: if the side to move has a pawn that threatens the other side's queen, then the pawn must capture the queen. The trouble with this rule is that in some cases the capture in question is illegal, as it would leave the king exposed. But it is certainly possible for people to adopt the rules of chess* anyway, presumably unaware of the conflict. In that case, there will eventually be a case in which a move is both required and prohibited.

Each of the examples just cited involves a kind of social convention, and so we have reasons for thinking that conventions can sometimes make inconsistent demands on their parties. If language is conventional in the same sense, then there should be a possibility of inconsistent rules or conventions of language as well. (The biggest difference is that in language, the terms of the convention are not given explicitly. But why should that matter?) In all cases of inconsistent rules, since one cannot actually both perform a given act and not perform it, some departure from the existing rules must take place. The "best" such departure is, arguably, to revise the rules and make them consistent. But this isn't always feasible (and pragmatically may not always be desirable), so the alternative is to simply muddle through and do whatever seems the most sensible. Either way, the response is inherently improvisational. It may be worth noting here that when presented with a case of the liar, most people do in fact just muddle through as best they can, in a way that seems to me to be improvisational rather than rule based. In any case, I don't think there is any inconsistency in the claim that a given system of obligations includes conflicts.

Another possible source of inconsistency for the present account is as follows. If the inconsistency theory is right, then speakers of English are committed to (a) the truth of the liar sentence (1), and (b) the non-truth of (1). That theory, moreover, is stated in English. Doesn't that mean the theory itself is committed to both the truth and the non-truth of (1)?

No, it doesn't. To see this, consider that while I did use English to state the inconsistency theory, in principle I needn't have. I could have stated the account in some other language—say, a consistent fragment of English. In that case, anyone who wants to assert the theory without also being committed to inconsistent sets of sentences need only confine herself to some consistent language in which the theory is statable. If this is possible—if there is a consistent language in which the inconsistency theory can be stated—then the act of asserting the theory need not be accompanied by any commitment to a contradiction, and therefore the theory itself does not imply any contradiction.

To put this point a bit differently, if the inconsistency theory is true, then we as speakers of English are committed to both the truth and the non-truth of (1). But this doesn't imply that the theory itself is committed to the truth and non-truth of (1). The theory takes no stand on that issue. As speakers of English, we may feel compelled to take some stand on the issue, and, indeed, as speakers of English we may be obligated to take conflicting stands on the issue. But it doesn't follow that the inconsistency theory itself takes any particular stand.

This all assumes that there is a consistent language—a consistent fragment of English, or otherwise—in which the inconsistency theory can be stated. If there isn't, then the inconsistency theory arguably becomes self-defeating or degenerates into dialetheism. This will be a problem if, and as far as I can see only if, the inconsistency theory requires the (ordinary) notion of truth for its formulation. Does it?

An old argument against inconsistency theories, due to Herzberger (1967), is as follows. Consider the claim that two sentences *A* and ~*A* are analytic. This will be the case

if *A* and *~A* are both logical consequences of some self-contradictory analytic sentence *B*, where *B* might be a contradictory instance of (T), for example. The classic definition of analyticity is as follows: a sentence is analytic if it is true by virtue of its meaning. In particular, an analytic sentence is *true*. But then we have that both *A* and *~A* are true. Furthermore, we presumably have that *~A* is true if and only if *A* is not true. In that case, we have shown that *A* is both true and not true. Thus, the claim that a sentence *B* is both analytic and contradictory is itself a contradictory claim. Finally, if the inconsistency theory is the claim that the instances of (T) are analytic, then by Herzberger's argument, the inconsistency theory is inconsistent.

In response, I never actually claimed that (T) is analytic, and more importantly, if I were to do so I certainly would not use the above definition of analyticity. In fact, I do think that "analytic" is an apt term for the special status of (T), but only if analyticity is understood in terms of language-generated commitments and not in terms of truth by virtue of meaning. As an aside, there's nothing sacred about the "true by virtue of meaning" definition of analyticity, which historically is only one of many.

A similar objection, also made by Herzberger, runs as follows. The inconsistency theory is a theory about the meaning of the word "true." Meaning is best understood in terms of truth conditions, or more generally of application conditions. But what, then, are the application conditions of the ordinary word "true"? That is, what is the extension of "true"? The answer cannot be: the unique extension that satisfies (T), since there is no such extension. There seems to be no way to explain (T)'s special status in truth-conditional or application-conditional terms.

I think it's pretty clear, then, that the inconsistency theory, while a theory of meaning, cannot be understood as a theory of anything resembling truth conditions. And this raises the broader question of how the present account fits into the more general study of language.

Truth conditional semantics, of course, represents just one approach to meaning. A theory based on inferential role semantics (as per Brandom (1994)) might accommodate the present account easily. Roughly speaking, inferential role semantics explains the meaning of an expression in terms of the inferences it participates in with respect to other expressions. The cases where inferential role semantics is most convincing are those of logical operators, with the associated inference rules providing the inferential role. The inconsistency theory of truth fits easily within this framework, provided the inferences can be inconsistent—and why can't they be? Moreover, the truth predicate strikes many as a logical operator, with the inferences from *A* to "*A* is true" and vice versa appearing to many (myself included) as logical inferences, suggesting that the truth predicate ought to be a good candidate for inferentialist treatment.

Of course, not everyone is an inferentialist, and indeed some sort of truth-conditional approach may be the most popular take on meaning. To those who are sympathetic to truth conditions (myself included!), I make the following suggestion. Facts about truth conditions must somehow supervene on facts about the use of language. How this takes place is not well understood, but may be thought of, roughly speaking, as involving a "fit" between the semantic facts and the use facts. Moreover, I suggest that these use facts should be understood as including normative facts, including facts about commitments to inferences. (These facts, in turn, must somehow supervene on still more basic facts, in a way that is not well understood but which might also be described as "fit.") Now in the case of an inconsistent predicate such as "true," the expected semantic fact—in this case, a fact about the extension of the predicate—is missing, because no possible extension of the predicate fits the use facts sufficiently. (Any such extension would have to obey (T), and none does.) We might describe this as a breakdown in the language mechanisms that normally produce referential facts. I would suggest that there are other, similar breakdowns in language, such as (some cases of) empty names. Be that as it may, while there isn't much useful we can say about the ordinary predicate "true" at the semantic level, we can still say something useful at the use level, namely, that there is a commitment to (T).

This is what I think we should say about inconsistent predicates in general, though there is a snag when the predicate in question is "true." Namely, on the account just sketched, the semantic facts include facts about reference and truth conditions. But if the use of "true" is governed by an inconsistent rule and lacks a proper extension, what sense does it make to talk about truth conditions at all? This is indeed a concern, but it assumes that the notion of truth that we use when talking about truth conditions is the same as the ordinary notion of truth that this paper is about. It need not be. In particular, I have been stressing all along the possibility of a revisionary notion of truth, and it may well be that one of the things we need a revisionary notion for is semantic theory. The feasibility of this project—i.e., of finding a paradox-free notion of truth that can be used in a semantic theory—is obviously an important question. Fortunately, there is a great deal of contemporary research devoted to this problem.

Let me end by describing two competing views of language. On one view, a language provides a mapping from sentences to propositions. Speakers can then use this mapping to commit themselves to various propositions by assertively uttering the corresponding sentences. Language determines what we can say, and only then do speakers decide what gets said. The language itself is transparent in that it doesn't impose any commitments or convey any information. In short, a speaker can opt into a language game without taking on any substantive commitments. I think this is a rather widespread and commonsensical view, but it is incompatible with the inconsistency theory. On that theory, speaking a natural language commits one to (T) and to (T)'s consequences, which are substantive. The medium and the message are less separate than the commonsense view suggests. This actually strikes me as a welcome conclusion—(T) is just one of many ways, I suspect, that the language we speak incorporates assumptions about the world we speak of—but it may also be one reason why the inconsistency theory is not more popular.

NOTES

1. Similar ideas were also expressed by Carnap (*Logical Syntax of Language*); see especially sec. 60. While the first systematic

development of the idea seems to be that of Chihara, the general notion of an inconsistency theory of truth was well known after Tarski's work, and there was sporadic discussion in the literature; see especially Herzberger ("Truth-Conditional Consistency").

2. Specifically, a set or relation is recursively enumerable iff it can be defined in the fragment of the language of arithmetic whose logical operators are &, v, ∃x, and ∀x<y. A set or relation is recursive (i.e., computable) iff it and its complement are recursively enumerable.
3. See, e.g., Priest, *Contradiction*.
4. It was suggested, for example, by Kripke ("Outline of a Theory of Truth"), and later defended in detail by Soames (*Understanding Truth*).
5. See Soames, *Understanding Truth*, for an account along these lines.
6. See Russell, *Mathematical Logic*; Parsons, *Liar Paradox*; and Burge, *Semantical Paradox*, among others.

BIBLIOGRAPHY

Barker, John. "The Inconsistency Theory of Truth" (Ph.D. diss.) Princeton University, 1999.

Brandom, Robert. *Making It Explicit*. Cambridge: Harvard University Press, 1994.

Burge, Tyler. "Semantical Paradox." *Journal of Philosophy* 76 (1979): 169–98.

Carnap, Rudolph. *The Logical Syntax of Language*. London: Kegan Paul, 1937.

Chihara, Charles. "The Semantic Paradoxes: A Diagnostic Investigation." *Philosophical Review* 88 (1979): 590–618.

Davidson, Donald. "Truth and Meaning." *Synthese* 17 (1967): 304–23.

Herzberger, Hans. "The Truth-Conditional Consistency of Natural Language." *Journal of Philosophy* 64 (1967): 29–35.

Hintikka, Jaakko. "Logica as a Theory of Computability." *APA Newsletter on Philosophy and Computers* 11, no. 1 (2011): 2–5.

Kripke, Saul. "Outline of a Theory of Truth." *Journal of Philosophy* 72 (1975): 690–716.

Parsons, Charles. "The Liar Paradox." *Journal of Philosophical Logic* 3 (1974): 381–412.

Priest, Graham. *In Contradiction*, 2nd ed. Oxford: Oxford University Press, 2006.

Russell, Bertrand. "Mathematical Logic as Based on the Theory of Types." *American Journal of Mathematics* 30 (1908): 222–62.

Soames, Scott. *Understanding Truth*. Oxford: Oxford University Press, 1999.

Tarski, Alfred. "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Logica* 1 (1935): 261–405. Translated as "The Concept of Truth in Formalized Languages," in *Logic, Semantics, Metamathematics*, edited by J. Woodger. Oxford: Oxford University Press, 1956.

Function Logic and the Theory of Computability

Jaakko Hintikka
BOSTON UNIVERSITY

ABSTRACT
An important link between model theory and proof theory is to construe a deductive disproof of S as an attempted construction of a countermodel to it. In the function logic outlined here, this idea is implemented in such a way that different kinds of individuals can be introduced into the countermodel in any order whatsoever. This imposes connections between the length of the branches of the tree that a disproof is and their number. If there are already *n* individuals in the countermodel that is being constructed, the next individual has to be considered in its relations to each of the *n* old ones, creating 2^n different cases and accordingly at least 2^n different branches. Hence a disproof procedure of a polynomial length is normally not equivalent with an exponential one. Because every computation can be represented as a deduction with the same number of constant terms, the same holds for nondeterministic computations. Apparent exceptions seem to come about if a branch created by a new individual *i* is redundant. But when the disproof is a shortest one (contains the minimum number of different constant terms) then not introducing that idle term at all would result in an even shorter disproof, violating the shortness assumption.

1. COMPUTATIONS AS DEDUCTIONS

The theory of recursive functions and computability was originally created in the context of logical problems, such as the *Entscheidungsproblem* for first-order logic.[1] Yet the precise relationships between logic and computation are not quite fully understood. In earlier papers I have pointed out how an arbitrary computation in a suitable first-order elementary arithmetic can be represented as formal first-order deduction using essentially the same number of constant terms.[2] To compute $f(x)$ for $x = a$ is then to deduce $f(a) = b$, which again corresponds to a formal disproof of $\neg(\exists y)(f(a) = y)$. The main reason why this is not trivially obvious is the quantificational structure of the ordinary first-order logic. In it, the arguments of Skolem functions must come from the quantifiers lower down in the same labeled tree. This imposes a simplified tree structure on the argument sets of those Skolem functions. This requirement is not satisfied by an arbitrary set of functions used in a set of equations defining a computable function.

This suggests using the proof-theoretical structure of deductions to explore the structure of arbitrary computations. Both processes are indeterministic computations with the same parameters (length and number of branches, etc.) characterizing both. In order to avoid the complications due to the quantificational labeled tree structure, it is advisable to build our discussion on a logic where quantificational relations are replaced by functional applications. Our first task is therefore to sketch a simple calculus which can be done by operating systematically on functions in first-order logic instead of predicates.

2. FROM PREDICATE LOGIC TO FUNCTION LOGIC

Functions have been the neglected stepchildren of first-order logic. Frege talks a lot about functions, but he makes serious use only of propositional functions. This is shown among other features of his thought by the fact that he would have avoided his entire puzzle about the cognitive value of identity statements if he had had functions among his nonlogical primitive identity statements. There cannot be any doubt about whether identities involving functions have cognitive information.

Likewise, Wittgenstein did not have functions among his nonlogical primitives in the *Tractatus*. If he had, he could have solved his color incompatibility problem by construing color as a mapping (function) from points in visual space into color space.[3]

It is often said that we can treat functions as relations of a special kind, that is, instead of a function $f(x)$ we could use a predicate $F(x,y)$ that applies whenever $f(x) = y$. This kind of selection of nonlogical primitives may perhaps be carried out in each given nonlogical theory, but it cannot be done in logic itself. The reason is that such a rewriting does not preserve logical properties. For each F used to replace f we would have to assume separately two things

(2.1) $(\forall x)(\exists y)F(x, y)$

(2.2) $(\forall x)(\forall y)(\forall z)((F(x, y) \ \& \ F(x, z) \supset (y = z))$

These are not logical truths about F. The *logic* of functions does not reduce to the *logic* of predicates. One cannot *logically* define a function in terms of predicates.

This holds a fortiori of constant functions, that is, of proper names of objects. They cannot be defined logically in purely descriptive terms. This logical truth is the gist of Kripke's criticism of descriptive theories of proper names.

If it is any consolation, in the other direction the semantical job of predicates can be done by functions, viz. their characteristic functions. If $P(x)$ is a predicate, we could change our language slightly and instead of $P(a)$ we could say $p(a) = d$ where d is a specially designated object and the characteristic function of P. This possibility of replacing predicates by functions in our logic is what is studied in this paper.

Hence, instead of any usual first-order predicate language (that includes =), we can use a language with only functions as nonlogical primitives. Naturally, we must also use the notion of identity expressed by =. The semantics for such a language can be assumed to be defined by means of the usual game-theoretical semantics.[4]

This paper is in the first place a survey of the fundamentals of such a function logic (of the first order), together with a couple of important applications.

For simplicity, it is in the following assumed once and for all that the formulas we are talking about are in a negation normal form. That is to say, the only connectives are ~, V, &, and all negation signs are prefixed to atomic formulas or identities.

A major simplification is immediately available, a simplification that is not available in predicate logic. Consider a formula of such a function language in its negation normal form. We can replace each existential formula $(\exists x)F[x]$ in the context

(2.3) $S[\text{—}(\exists x)F[x]\text{—}]$

without any change of the intended meaning

(2.4) $S[\text{—}F[f(y_1, y_2 \ldots c_1, c_2, \ldots)]\text{—}]$

where is a new function called a Skolem function of $(\exists x)$ and $(Q_1 y_1)(Q_2 y_2)$ ... are all the quantifiers on which $(\exists x)$ depends in S. Moreover, c_1, c_2, ... are all the constant terms on which $(\exists x)$ depends in that context. After the change, the function f now does the same job in (2.4) as the quantifier $(\exists x)$ in (2.3).

The result is a language in which there are no existential quantifiers and in which all atomic expressions are negated or unnegated identities. Such a language is here called a function language and its logic a function logic.

What are they like? Such a logic is a kind of general algebra. All logical operations on formulas, including application of rules of inference, are manipulations of identities by means of substitutions of constant terms for universally bound variables, plus the substitutivity of identity and propositional rules. The only quantifier rule needed is the substitution of a term for a universally quantified variable. The rules for existential quantifiers are taken care of by treating their Skolem functions just like any other functions.

This paper is an exploratory study of function languages.

What are they like? Logical operations, including formal proofs, often become much simpler when conducted in a function language. This is especially conspicuous in theories like group theory where it is much more practical to express axioms in terms of functions and equations involving functions than by means of quantifiers.

In the elimination of existential quantifiers in terms of Skolem functions, the notion of dependence was used, both for dependencies of quantifiers on other quantifiers and for dependencies on constants. Here the semantical meaning of the dependence of a quantifier $(Q_2 y)$ on another quantifier $(Q_1 x)$ means the ordinary ("material") dependence of the variable y on the variable x. In traditional first-order logic this is expressed by the fact that $(Q_2 y)$ occurs within the scope of $(Q_1 x)$. In the Skolem representation such dependence amounts to the fact that x occurs among the arguments of the Skolem function associated with $(Q_2 y)$. The dependence of $(Q_2 y)$ on a constant c is likewise expressed by c's occurring as an argument of the Skolem function replacing $(Q_2 y)$.

3. SKOLEM FUNCTIONS AND SCOPE

All the same modes of reasoning can be represented in function logic as can be represented in the usual first-order predicate logic.

Function languages and function logics can be defined in their own right by specifying the functions that serve as its primitives, without any reference to a paraphrase from an ordinary first-order predicate language. For instance, since Skolem functions behave like any other functions, they do not need any existential quantifiers to paraphrase. Such function languages are in fact logically richer than ordinary first-order predicate languages. The reason is the fundamental fact that not all sentences of a function language can come from a predicate language expression.[5] This reason is worth spelling out carefully. The key fact is the tree structure of predicate language formulas created by the scopes of quantifiers and connectives. These scopes are indicated by pairs of parentheses. In the received first-order logic, these scopes are nested, which creates the tree structure, that is, a partial ordering in which all branches (descending chains) are linearly ordered.

Since dependence relations between quantifiers and connectives are indicated by the nesting of scopes, these dependence relations also form a tree. Depending on

precisely what kind of logic we are dealing with, certain scopes are irrelevant to dependence. In this paper, like in the usual IF (independence friendly) logic, only dependences of existential quantifiers on universal ones are considered. (But see below for more details.) The arguments of a Skolem function come from quantifiers and constants lower down in the same branch, as one can see from (2.4). Hence, the argument sets of Skolem functions must have the same tree structure as the formulas they come from, suitably reduced. There is no reason why the argument sets of the functions in a function language formula or set of formulas that do the job of existential quantifiers should do so. Hence, a function logic is formally richer than the corresponding predicate logic. It turns out that this also makes it much richer semantically.

Indeed, as is spelled out in Hintikka (2011a), this tree structure restriction nevertheless holds only for languages using the received first-order predicate logic. A subset of $\{y_1, y_2 \ldots c_1, c_2, \ldots\}$ can be the argument set of the f in (2.4). Hence, the function logic we are dealing with here is richer than ordinary first-order logic. If the only extra independences allowed are independences of existential quantifiers of universal ones, the resulting logic is equivalent to the usual IF logic as explained in Hintikka and Symons (forthcoming) and later in this paper. An independence-friendly (IF) first-order language is not expressionally poorer with respect to quantifiers than the corresponding function language. In such a predicate language, any subset of $\{y_1, y_2 \ldots c_1, c_2, \ldots\}$ can be the argument set of the f in (2.4), according to which quantifiers and/or constants outside the quantifier $(\exists x)$ depends on.

Already at this point we see that the step from predicate languages to function languages strengthens our logic greatly and in fact throws light on one of the most important logico-mathematical principles. In this step the job of existential quantifiers is taken over—naturally, indeed inevitably and unproblematically—by Skolem functions. (On a closer analysis, this unproblematic character of Skolem functions in this role is based on their nature as the truth-makers of quantificational sentences.) But the existence of all these Skolem functions has the same effect as the assumption of an unlimited form of the so-called "axiom" of choice. This mathematical assumption thus turns out to be nothing more and nothing less than a valid first-order logical principle, automatically incorporated in function logic.[6]

In other ways, too, the apparently unproblematic step from predicate logic to function logic brings out the open fundamental questions. One of the interesting features of function logic is that we can by its means express the same things that are in IF logic expressed by means of the independence indicator slash /. In order to see how this is done, it may be pointed out that many of the limitations of ordinary first-order logic are due to the fact that the notion of scope is in it overworked.[7] Semantically speaking, it tries to express two or perhaps three things at the same time. The first two may be called the government scope and binding scope. The distinction between the two is obviously the same as Chomsky's distinction between his two eponymous relations, although Chomsky does not discuss their semantical meaning.[8]

Government scope is calculated to express the logical priority of the different logical notions. In game-theoretical semantics, it helps to define the game tree, that is, the structure of possible moves in a semantical game. The nesting of government scopes must hence form a tree structure. It is naturally expressed by parentheses. In function logic, such parentheses are needed mainly for propositional connectives. The only quantifiers are universal ones, and as long as we can assume (as is done in ordinary first-order logic and in the simpler form of IF logic) that universal quantifiers are independent of each other and of existential quantifiers, their binding scope does not need to be indicated by parentheses as long as different variables are used in different quantifiers. For the justification of this statement, see sec. 4 below.

Formal binding scopes are supposed to indicate the segment of a sentence (or formula or maybe discourse) in which a variable bound to the quantifier is grammatically speaking an anaphoric relation. There is no general reason to expect that such a binding scope should be a connected part of a formula immediately following a quantifier, even though that is required in the received first-order logic. There is no such requirement in the semantics of natural language.

Such binding is automatically expressed in a formal language by the identity of the actively used variables. All we have to do is to require that different quantifiers have different variables.

However, this leaves unexpressed a third kind of important relation of dependence and independence, over and above the dependence and independence of quantifiers and constants. It is the dependence and independence of other notions, such as connectives. As long as we can assume that these dependencies are so simple that the semantical games we need are games of perfect information, those dependence relations are captured by the nesting of government scope. But this assumption has turned out to be unrealistically restrictive in formal as well as natural language.

In order to overcome this restriction, in the usual form of IF logic there is an independence indicating symbol, the slash / that overrules the government scope as an (in)dependence indicator. Do we need it in function logic? In function logic, we have a different way of indicating the dependence of a quantifier on others and on constants. The only quantifiers we are using are existential ones, represented by Skolem functions plus sentence-initial universal quantifiers. The dependence of an existential quantifier $(\exists x)$ on $(\forall y)$ is to have y among the arguments of its Skolem functions and likewise for constants.

In any case in a function logic all quantifier dependencies and independencies as well as dependence relations between quantifiers and constants can be expressed without any explicit independence indicator.

4. QUANTIFIER-CONNECTIVE (IN)DEPENDENCIES

One more class of dependence and independence phenomena is nevertheless constituted by the relations of quantifiers and connectives to each other. From game-theoretical semantics it is seen that the question of informational dependence or independence automatically arises also in the case of application of quantifier rules and of

rules for connectives. Somewhat surprisingly, an examination of these relations leads to serious previously unexamined criticisms of the traditional first-order predicate logic and of Tarski-type truth definitions.[9]

These criticisms are best understood by means of examples. Consider for the purpose a sentence of the form

(4.1) $(\exists x)(A(x) \supset (\forall y)A(y))$.

This is equivalent with

(4.2) $(\exists x)(\sim A(x) \vee (\forall y)A(y))$.

This (4.1) can be considered as a translation of an ordinary discourse sentence.

(4.3) There is someone such that if she loses money in the stock market next year, everyone will do so.

This is obviously intended to be construed as a contingent statement, and hence cannot be interpreted so as to be logically true. Yet (4.1) and (4.2) are logically true if a Tarski-type truth definition is used. For there exists a truth-making choice $x = b$ no matter what possible scenario (play) is realized, that is, independently of which choice satisfies the disjunction

(4.4) $\sim A(x) \vee (\forall y)A(y)$.

There are two possibilities concerning the scenario that is actually realized: Either (i) everybody loses money or (ii) someone does not. In case (i) any choice of $x = b$ satisfies (4.4). Then b must lose his money along with everybody else.

If (ii), the someone (say d) does not lose and can serve as the choice $x = d$ that satisfies (4.4). Accordingly, truth-making choices are always possible. Hence, on a Tarski-type truth definition (4.1)–(4.3) must be true in any case in any model; in other words, they must be logically true.

However, b cannot be the same individual as d, for the two have different properties. Hence, there need not exist any single choice of x that satisfies (4.4) no matter how the play of the game turns out, which obviously is the intended force of (4.3). What happens is that on the intended meaning of (4.3), the choice of $x = b$ or $x = d$ is assumed to be made without knowing what will happen to the market, that is to say, independently of which scenario will be realized. In terms of semantical games, this means that the choice of the disjunct in (4.2) or (4.4) cannot have been anticipated in the choice of the individual (b or d). In logical terms, this means that the existential quantifier and the disjunction are independent of each other. This independence is implemented by replacing the disjunction V in (4.2) by $(V/\exists x)$.

The general issue is the relationship between formulas of the form

(4.5) $(\exists x)A[x] \vee B[y]$ and

(4.6) $(\exists x)(A[x] \vee B[y])$

as well as between

(4.7) $(\forall x)A[x] \,\&\, B[y]$ and

(4.8) $(\forall x)(A[x] \,\&\, B[y])$.

i.e., where x does not occur in $B[y]$: Here the equivalence of (4.7) and (4.8) is what justifies us to move all universal quantifiers in a function logic formula into its beginning.

If we do not have the independence-indicating slash / at our disposal, we have to assume an interpretation (a semantics) of first-order expression like (4.1)–(4.2) different from the conventional ones. This conventional semantics is a Tarski-type one. It does make the two equivalences valid, but it violates the intended meaning of our informal as well as formal expressions. In other words, a Tarski-type semantics is an inaccurate representation of the intended meanings of sentences like (4.3) and of their usual slash-free formal representations.[10]

In contrast, GTS yields the right reading, but only when we assume an independence between $(\exists x)$ and V in (4.1)–(4.2). Our function logic does not include separate independence indicators, wherefore we have to assume the independence in question throughout.

A proof of logical truth is a kind of reversed mirror image of semantic games. In such a proof, we are trying to construct a model in which the formula to be proved is false. The independence of the kind just pointed out means in effect that all the alternative models that we may have to contemplate in the construction must have the same domain of individuals. This shows that the same independence assumption is tacitly made also in normal mathematical reasoning.

As to the rest of the semantic of our function logic, negation ~ is supposed to be defined in the usual game-theoretical way (exchange of the roles of the verifier and the falsifier), which means that it is the strong dual negation. The contradictory negation ¬ is interpreted game-theoretically only on a sentence-initial position or else prefixed to an identity.

5. FORMATION RULES

Thus, function logic exhibits several interesting novelties even though it was originally introduced as little more than a paraphrase of the familiar predicate logic in terms of functions instead of predicates. Formally, our function logic nevertheless seems to be quite straight-forward. For one thing, we can formulate the formation rules for function calculus without using independence indicators, or any other symbols. They can be expressed as follows.

The nonlogical primitive symbols are functions f, g, h, ... of one or more argument places, individual variables x, y, z, ..., the universal quantifiers $(\forall x)$, $(\forall y)$, $(\forall z)$ (please note that they do not come with parentheses trailing them), …. , plus primitive constants a, b, c, ...

The primitive logical symbols are ~, &, V, = plus Skolem functions with one or more argument places s, t, u, ...

A term is defined in the usual way.

(i) A primitive constant or a variable is a term.

(ii) If f is a function with k argument places and t_1, t_2, ..., t_k are terms, then so is $f(t_1, t_2, ..., t_k)$.

(iii) The same for Skolem functions.

A term without variables is a constant term.

The rules for formulas are simple:

(i) If t_1, and t_2 are terms, $(t_1 = t_2)$ is a formula (an identity).

(ii) Negations of identities $\sim(t_1 = t_2)$ (abbreviated $(t_1 \neq t_2)$ are formulas.

(iii) Truth functions in terms of & and V of formulas are formulas

We will take $(F_1 \supset F_2)$ to be the same as $(\sim F \vee F_2)$.

(iv) If F is a formula containing free occurrences of a variable x, then $(\forall x)F$ is a formula.

The variable x in $(\forall x)F$ is said to be bound to $(\forall x)$, otherwise free.

A formula so defined is always in a negation normal form in which all negations are negations of identities.

A couple of important general explanations are still in order. The general theoretical interest and its usefulness for applications of function logic lies in the fact that it captures much of the force of IF logic without apparently going beyond the resources of ordinary first-order logic. This means two things: (a) not using any special independence indicators and (b) using overtly no negation other than the one defined by the rules of the semantical games.

As far as (i) is concerned, it is easily seen what happens. The job of expressing dependencies and independencies between variables is in function logic taken over by Skolem functions. Using them in dependence of a variable x can be expressed by leaving x out from the arguments of a Skolem function.

The semantical stipulations above make the following pairs of formulas equivalent and hence interchangeable:

(5.1) $(\forall x)A[x] \,\&\, B$

$(\forall x)(A[x] \,\&\, B)$

(5.2) $(\forall x)\, A[x] \vee B$

$(\forall x)(A[x] \vee B)$

It is assumed, as the notation shows, that x does not occur free in B. This means that each formula has a normal form in which it has the form of a truth-function of identities governed by a string of universal quantifiers. All logical operations are substitutions of terms for universal quantifiers and applications of the substitutivity of identicals. This illustrates further the role of function logic as a kind of universal algebra.

Indeed, function logic throws interesting light on the very notion of universal algebra, especially on its relation to logic and on its status as a codification of symbolic computation in analogy with numerical computation.[11]

6. RULES OF PROOF

Likewise, the formal rules of proof, or rather disproof, are obtained in a straightforward way from the corresponding rules for predicate logic, and so is their semantical (model-theoretical) meaning. Semantically—and hence intuitively—speaking, a sentence in a function language can be thought of as a recipe for constructing a description of a scenario (world) in which it would be true. Hence, the primary question about its logical status is whether the description is consistent, in other words whether is satisfiable. If not, is logically false (inconsistent). This can be tested by trying to construct a description of a model in which would be true. Such a construction will take the form of building step by step a set of formulas which is obviously consistent. Model sets in the usual sense are known to be so.[12]

A disjunction splits such a model set construction into branches. If all of them lead to contradiction, S is inconsistent; if not, S is satisfiable.

The explicit rules for proof are variations of the corresponding rules for predicate logic disproofs. They take the form of rules for constructing a model set for a given initial formula or set of formulas. The construction can be divided into different branches.

The propositional rules are the same as in predicate logic.

(R.&) If $(F_1 \,\&\, F_2) \in B$, add F_1 and F_2 to B

(R.) If $(F_1 \vee F_2) \in B$, divide the branch into two, B_1 and B_2 with $F_1 \in B_1$ and $F_2 \in B_2$.

Likewise, the rule for identity is the same.

(R.=) Substitutivity of identity

Since existential quantifiers have been eliminated in terms of Skolem functions, no rules are needed for them.

The counterpart to the predicate logic rule for universal quantifiers is the following:

(R.A) If $(\forall x)\, F[x] \in B$ and if the constant term t can be built out of functions and constants occurring in (the members of) B, then $F[t]$ may be added to B.

In these rules, B is the initial segment of a branch so far reached in the construction. From what was found earlier in section 4, it is seen that the restriction on t can be somewhat relaxed. It was shown there that in the kind of logic that deals with a fixed domain, quantifiers and disjunctions are independent of each other. This corresponds in function logic to allowing in (R.A) as the substitution value of t any term that is formed from functions and constants in any initial segment B of any branch so far reached, and not just in B.

And this obviously means allowing as t any constant term formed out of the given functions and constants of the initial S plus the Skolem functions of S. The rule (R.A) thus emended is called (R.A)* The rules the, formulated are (R.&), (R.v), (R.=) and (R.A)*.

We need a rule for negation. Since we are dealing with formulas in a negation normal form, all negations occur in prefixes of identities, it suffices to require the obvious:

(R.~) A branch B is inconsistent if $F \in B$, $\sim F \in B$ for any F, or $\sim(t = t) \in B$ for any term t.

A moment's thought shows why the prohibition against ($t \neq t$) is enough to take care of identities. For by substitutivity of identity $\sim(t_1 = t_2)$ and $(t_1 = t_2)$ it follows that $\sim(t_1 = t_1)$.

We can formulate an equivalent proof (attempted model construction) method. It will be called the internal construction method. It takes the form not of building a set of formulas starting from S, but of modifying S step by step from $S_0 = S$ to S_1, S_2,... . Different initial segments of branches of the disproof construction then become different maximal parts of the single formula S_i under consideration not separated by V and secondarily lists of subformulas in them. In other words, we can join different branches of an attempted proof tree as disjuncts so as to become parts of a single formula separated by V (after the members of the same branch are combined into conjunction). The construction of the sequence S_1, S_2,... proceeds according to the rules (r.A) and (r.=).

(r.A) If $(\forall x)\, F[x]$ is a subformula of S_i, replace it by $(\forall x F[x] \,\&\, F[t])$

Here t can be any constant term formed from the given constants and functions of S plus the Skolem functions of S. This rule can be generalized by allowing the substitution-value term contain variables universally bound to a quantifier (in the context in which $(\forall x)F[x]$ occurs). This extension can easily be seen not to widen the range of formulas that can be proved.

If we had not made connectives and quantifiers independent of each other, we would have to require that the Skolem functions in t occur in the same branch.

No rule for conjunction is needed. The negation rule can be formulated in the same way as before, but taking the notion of branch in the new sense.

If quantifiers and connectives are not made independent of each other as explained above, a new constant term may be introduced only if all its functions and constants already occur in the same branch. This rule can be generalized by allowing the substitution-value term to contain variables universally bound to a quantifier (in the context in which $(\forall x)$ $F[x]$ occurs). This extension can be easily seen not to widen the range of formulas that can be disproved.

If we had not made connectives and quantifiers independent of each other, we would have had to require that the Skolem functions in occur in the same branch.

We also need a suitable rule of the substitutivity of identicals:

(r.=) If $(t_1 = t_2)$ is a subformula of S_i and A is a subformula in the same branch as $(t_1 = t_2)$, then the A can be replaced by $(A \,\&\, B)$, where B is like that t_1 and t_2 have been interchanged in some of their occurrences.

Thus, a construction of a branch of a proof tree in search of a model set is literally the same as is a construction of a branch in the expansion of the given initial sentence that is being tested for consistency. The rules were just listed.

In either version the proof construction, serving as taking the form of a disproof method, is easily seen to be semantically complete.

The two equivalent proof methods will be called external and internal proofs.

From the semantical perspective, an attempted proof of S is an attempt to construct a model or strictly speaking a model set for it. The rules (R) and (r) regulate the introduction of new individuals into the model construction. It is to be noted that model-theoretically (semantically) speaking, a single application of the rule (R) can in effect introduce several new individuals at the same time. This is because of the nesting of terms. A complex term may contain as an argument a likewise complex (albeit simpler) term. In keeping track of the number of individuals introduced into an experimental model construction, all different constituents of constant constituent terms must be counted.

If quantifiers and connectives are not made independent of each other as explained above, a new constant term may be introduced only if all its functions and constants already occur in the same branch.

If it is required that new terms are introduced one by one, we can simply allow only the introduction of terms that are not nested. However, then we have to allow the introduction of terms that are not constant but contain (universally bound) variables. As was noted, this extension of our rules is obviously possible.

7. ON THE STRUCTURE OF FUNCTION LOGIC

In all their simplicity, these sets of rules of proof are remarkable in more than one way. In the internal method, there are no restrictions as to when rules are applied, except of course for the presence of the subformula to which a rule is applied. In particular, since the universal quantifiers remain the same throughout a proof, any constant term can be introduced at any time. The order of their introduction is completely free.

This throws some light on the nature of the entire proof theory. As proof theory for first-order theories is usually developed, a great deal of attention and care has to be expended on questions concerning the order and possible commutability of the rules. We can now see that much of such a problematic is caused from our perspective by unnecessary restrictions on the proof rules. For one typical thing, in the usual treatments of first-order predicate logic existential instantiation can be performed only on a sentence-initial existential quantifier. If so, in each the new term $f(t_1, t_2, \ldots)$, f must be the Skolem

function of a sentence-initial existential quantifier and t_1, t_2, ... constant terms previously introduced.

If we assume that all our formulas are sentences, a simple inductive argument using induction on the complexity of the given constant term shows that any constant term can be formed in accordance with this restriction by repeated application of restricted introductions. We only need to proceed from the outside in the introduction of new constant terms $f(t_1, t_2, \ldots)$ where f is a Skolem function. Hence the restriction does not make any difference to the class of provable formulas. This means in turn that what can be proved by the usual methods, for instance by means of the familiar tree method.[13] Since these methods are known to be complete, we obtain as a by-product a verification of the completeness of the set of our rules of proof.

In general, the flexibility of our proof rules allows us to see what in formal proofs is essential and inessential and thereby to have an overview of their structure. This structure involves two main elements, on the one hand the branches one by one with their properties, most prominently their length, and on the other hand the tree structure formed by the branches collectively, especially the number of branches.

The length of a proof branch can be measured in different ways. The number of formulas in a branch or in an initial segment of a branch B is called its overall length l. The number of different constant terms in B is called its combinatorial length c. The two are connected with each other. Obviously $l > c$. Also, from a given finite branch with n individuals $p(n)$ of new formulas can be formed without universal instantiation that are not propositionally equivalent with any of the old ones. Hence, the combinatiorial versus overall distinction becomes redundant when applied to deductive problems.

Assume that in B we have a number of functions with the total number of n argument-places. This number is determined by the structure of the input formula S, and hence independent of c. Out of them we can form at most c^n equations wherefore $l \leq c^n$. This is a polynomial function of c.

Function logic offers an overview of possible proofs of S. All that is needed is the finite list of substitutions of a constant term t for a universally quantified variable y that turns S into contradiction. Here t is formed from constants and constant functions of S plus from its Skolem functions. Each of these constant terms (mostly function terms) shows how the individual it represents is built up, and together they thus specify the structure of the proof. This structure is crucial both strategically (heuristically) and from the stand-point of strictness of proof. These are among the heuristic features of function logic proofs that create an analogy between proof search and experimental model construction.

8. LENGTH VERSUS NUMBER OF BRANCHES

One of the most important questions here is: How does the length of branches influence the number of branches? Let us assume that we are given a disproof of a given formula with the shortest combinational length. Let us consider the last step in the longest branch of the disproof that introduced a new constant term t_c into the disproof. In order to contribute to the argument, it must convey information about how the new individual is related to the old ones. Whatever the syntactical form of the new formula is, it has the same force as

(8.1) $\quad \bigwedge_i \bigvee_{ij} A_{ij} [t_c, t_{ij}]$

Here $i=1,2,\ldots,c-1$ while $A_{ij} [t_c, t_{ij}]$ are (negated or unnegated) identities. If all disjunctions have at least two disjuncts, the construction (disproof) splits into at least 2^{c-1} branches.

How can we exclude this possibility? Suppose that one of the disjunctions is reduced to a single formula $A_{ik} [t_c, t_k]$ where $1 \leq k \leq c-1$. Then the relationship of t_k to t_c is determined by the formula before the last introduction. By letting t_m play the role of t_c we obtain

(8.2) $\quad \bigwedge_i \bigvee_{ij} A_{ij} [t_m, t_{ij}]$

From (8.2) it can be seen that the relationship of t_k to any old individual t_m is likewise determined. In the disproof, it therefore does not need to be instantiated in order to complete the disproof. This would violate the assumption that we are dealing with the shortest disproof.

Hence the branch under scrutiny separates into at least 2^{c-1} branches. In other words, the number of branches in a disproof is an exponential fraction of their length.

In the light of the semantical idea of model construction this result is so intuitive as to be nearly obvious. Whenever a new individual is introduced into the construction, we have to examine its relationship to each of the $c-1$ earlier ones. If none of these examinations is redundant, there will be at least two different possibilities to be considered, and hence at least 2^{c-1} branches.

This result is an important example of the way we can reach insights into the structure of function logic proofs. We will call it the Branching Theorem. It can of course be formulated by reference to the structure of predicate logic proofs, but the initial absence of Skolem functions makes a demonstration of the theorem much clumsier.[14]

The Branching Theorem illustrates the ways in which the disproof theory of function logic throws light on the strategic problems of logical proofs. The insights it offers are made vivid by our interpretational perspective on disproofs as model constructions. The methodology that emerges is (for good reasons) reminiscent of the usual problem situation in traditional axiomatic geometry. In order to prove a theorem, one first drew a figure to "illustrate" it. (Logically speaking, one used instantiation rules in order to avoid quantificational reasoning.) As every student soon learned, the main difficulty in theorem proving was how to find the right "auxiliary constructions," to introduce the right geometrical objects. After their introduction, the proof was almost obvious, turning typically on examining the relationships between the different parts of the resulting figure. This strategic importance of "auxiliary constructions" is an example of the proof-theoretical significance of the selection of the constant terms that, intuitively speaking, introduce new individuals.

The proof theory of function logic turns out to illuminate in a similar way the strategic aspects of computation and semantical theorem proving.

9. THE PLACE OF FUNCTION LOGIC IN THE HIERARCHY OF LOGICS

Function logic is thus a logic of its own, conforming to what is in our days referred to as a "logical system," that is, an axiom system with a finite number of formal rules. This formal system has two somewhat different interpretations (semantics) depending on whether the single negation it uses is taken to be the contradictory negation ¬ or the dual (strong) negation ~.

How is this system on either reading related to other logics? Among the objects of such comparison there are (i) the received first-order logic, and (ii) IF logic in the usual narrow sense in which existential quantifiers can be independent of universal ones even though they are within their formal scope. If we also allow sentence-initial contradictory negation ¬, we have an extended IF logic (EIF logic) with two different negations. It has two symmetrical halves of which IF logic is one. The other one is a mirror image of IF logic, the logic of the contradictory negations of IF sentences. With an above justification, one might be tempted to suggest that the two halves of EIF logic are "really" one and the same logic looked upon from two different directions. It appears that EIF logic is the logic mathematicians use as their basic logic.

These halves can be compared with each other and with function logic by seeing what they correspond to in a more traditional hierarchy of logics. IF logic is equivalent with $\sum^1_1$ fragment of second-order logic, that is the logic of sentences of the form

(9.1) $(\exists X_1)(\exists X_2) \ldots F[X_1, X_2, \ldots]$

where X_1, X_2, are second-order variables (function or predicate variables) and F is an ordinary first-order formula (without slashes). The other half of EIF logic corresponds to $\prod^1_1$ fragment of second-order logic, that is, logic of sentences of the form

(9.2) $(\forall X_1)(\forall X_2) \ldots F[X_1, X_2, \ldots]$

Now function logic was introduced so as to be equivalent with ordinary first-order logic, apart from the unavoidable modification explained in section 4 above. This is true, but it requires an explanation. Both logics are one-negation logics, but this single negation can in principle be interpreted either as ¬ or as ~.

From the validity of the equivalence of (5.1) it follows that each sentence of function logic can be brought to an equivalent form in which all universal quantifiers are in the beginning of a formula and all existential quantifiers have been eliminated in favor of Skolem functions, in other words in the form

(9.3) $(\forall y_1)(\forall y_2) \ldots F[y_1, y_2, \ldots]$

where F is a complex IF formula. It is IF because it may contain ~. Now this is equivalent with a formula of the form

(9.4) $\neg\exists(y_1)\exists(y_2) \ldots \neg F[y_1, y_2, \ldots]$

which is the mirror image of an IF formula. This means that function logic is in effect the other half of EIF logic, the familiar half being the IF logic. This shows the location of function logic in the hierarchy of different logics.

It also throws interesting light on function logic itself as well as on its twin, IF logic. It would be tempting for a theorist of different logics to dismiss function logic as a mere mirror image of IF logic. However, it turns out that function logic considered independently can illuminate the nature and status of IF logic.

For one thing, some philosophers have criticized IF logic because it does not allow for a complete formal proof procedure. This could be shown to be a misguided requirement in any case, and here it can be noted that function logic does have a complete proof procedure, for instance the one outlined in section 6 of this paper. It would be absurd to maintain that IF logic is not a genuine logic while function is. Their entanglement with each other is illustrated by the fact that while IF logic does not have a complete formal proof procedure, it has a complete *dis*proof procedure, that is, recursive enumeration of logically false formulas. The mirror relation shows up in the fact that function logic does not have a complete disproof procedure.

Even though function logic and IF logic are closely intertwined, we can see that they are each worth formulating on their own, among other things, to serve different applications. It was pointed out earlier that function-theoretical ideas can be used to enrich proof theory. Function logic can be seen as one possible realization of the old idea of an universal algebra.[15] In the rest of this paper function logic is used to study the theory of computation. In contrast, in its usual form IF logic is a useful tool in analyzing higher-order and set-theoretical reasoning. This role of function logic as a virtual theory of computability shows that it cannot be dismissed as being merely a mirror image of the usual IF logic. It is a way of making a proof theory applicable to the theory of computability, at the same time as it clarifies and simplifies the structure of proof theory itself.

10. ON NUMERICAL COMPUTATION

Function logic thus throws interesting light on the nature of different issues on the first-order logic level, including the structure of logical proofs. The ultimate aim of this paper is broader, however. This is also calculated to discuss the nature of computability and the structure of computations. This is obviously too broad a subject to be dealt with exhaustively in one paper. But what is it that connects function logic with the general theory of computability?

Computability is an elusive topic that is not approached most effectively by starting from an attempted definition. A more general concept is what we propose to call here algorithmic computability. It proceeds from a finite set of AX of equations in terms of universally bound variables and constants and proceeds by means of the two rules of (i) substitutivity of identicals and (ii) substitution of constants for universal variables. We can assume that the axioms in AX are formulated in the language of a suitable elementary arithmetic without quantifiers. The members of AX are equations of function terms or negations of such equations. These equations are built out of a number of auxiliary functions g_1, g_2, … plus constants in addition to the function to be computed.

Actually, this is not general enough even though it is a good approximation of what actually happens in different terms of computability. Normally, instead of a set of equations or, equivalently, their conjunction, AX also typically includes some disjunctions of equations. Hence it may also contain conditionals and biconditionals of equations. For example, in primitive recursive computation we have among our starting-points besides plain equations also a biconditional and the negation of an equation (see below).

As a consequence of this extension, we need also propositional inferences beside the two principles (i)–(ii).

We will discuss computability in the form of arithmetical computability. In order to do so we need a fragment of elementary arithmetic, in the first place, the basic equations for primitive recursive functions. They can be formulated as follows.

In order to be able to speak of numbers and numerals, some number theory must be assumed to be given. Among other things, it is assumed that the usual primitive recursive equations are included in AX. Then the functions g will include the successor function $s(x)$, addition $x + y$, and multiplication $x \cdot y$. The equations are the obvious ones

(10.1) $s(x) \neq 0$ $\quad (x = z) \leftrightarrow (s(x) = s(z))$

(10.2) $(0 + x) = x$ $\quad s(y) + x = s(y + x)$

(10.3) $(x \cdot y) = 0$ $\quad s(y) \cdot x = (y \cdot x) + x$

These constitute the algorithmic basis for primitive recursive computability. General recursive computability of (partial) functions can likewise be that contains in effect (10.1)–(10.3), definitions of certain merely formal functions (projection, composition of functions) plus whenever needed, the equations for minimization of functions. These minimization functions take us from a given (computable) function to $g(x, y)$ to a new one f. This is accomplished by the assumption to be included in AX.

(10.4) $(\forall x){\sim}g(x, y) \vee \min(x, f(y)) = f(y)$

Here min can be defined by

(10.5) $\min(0, y)$ $\quad \min(1,0) = 0$

$\min(s(x), s(y)) = s(\min(x, y))$

The sets of equations that can serve to define general recursive (partial) functions consists of the equations for primitive recursive functions (10.1)–(10.4) and definitions of minimization functions (10.5), known to equal the usual sense of arithmetical computability. This sense of computability equals in turn among other things, Turing machine computability.

What is it that makes this a natural explication of computability in general? We are not trying to answer this question here. We can nevertheless define a wide sense of the computability of a partial function f. Assume that we are given any consistent "algorithm" AX, which consists of a number of equations that involve the function f to be calculated, function from (10.1)–(10.5) and a number of auxiliary functions $g_1, g_2, \ldots$. Consider now a proposition of the form

(10.6) $(\forall x){\sim}(f(n) = x)$

where n is a numeral. If the conjunction of (10.6) and AX is inconsistent, this inconsistency is provable in function logic. Such a proof produces a numeral m such that

(10.7) $\neg{\sim}(f(n) = m)$

is provable in function logic.

Assuming consistency, this means that either $f(n) = m$ is true and m is the computed value of $f(n)$ or else that $f(n)$ is undetermined. For in the latter case, any equation of the form $f(n) = h$ where $h \neq m$ contradicts $f(n) = m$. (A function cannot have two different values for the same argument.)

In this way AX defines a partial function that is in a natural sense computable. We will call such a function functionally computable. All general recursive functions are functionally computable, but the inverse relation remains to be studied, as do most properties of functional computability.

11. FUNCTION LOGIC AS A LOGIC OF COMPUTABILITY

We are now beginning to see the bridge that leads from function logic to a theory of computability. Or rather we can now see that no bridge is needed, for the two already are virtually the same theory. The deductive rules on which function logic is based are the very same rules of computation on which a theory of computation can be built, viz. the two substitution rules plus a modification of propositional logic. A computation of $f(n)$ on the basis of the algorithm codified in AX is the same process as the deduction of contradiction in function logic from

(11.1) AX & $(\forall x){\sim}(f(n) = x)$

Hence the study of function logic *ipso facto* is a theory of computability. In other words, different general questions about computability and algorithms are equivalent with questions about function logic. The job algorithms do becomes that of deductive premises. Accordingly, algorithms can be studied and compared with each other in the same way as deductive premises. Among other problems various questions concerning the correctness of algorithms can be studied in a simple problem situation. The correctness of an algorithm for a function $f(x)$ equals essentially the question of the arithmetical truth of the

(11.2) $(\exists g_1)(\exists g_2) \ldots \mathrm{AX}[f, g_1, g_2, \ldots]$

Likewise, steps in in a computation correspond to steps in a deductive reasoning. As a consequence structures of computation correspond to structures of deductions and are virtually identical with them. Accordingly, results concerning the former automatically become results of the latter.

The most interesting application of these observations is that from the Branching Theorem we can see that P≠NP, in other

words solve the P vs. NP problem.[16] At first sight, the two deal with different subject matters. The P vs. NP problem is whether the job of any polynomial length indeterministic computation procedure can be done by polynomial length deterministic procedure. The connection to function logic is created by the fact that any computation of was seen to be tantamount to a disproof of $(\forall x)\sim(f(n) = x)$ from a relevant AX. In particular, the same constant terms are used in the two processes—in so far as they are at bottom the same process.

Attempted disproof of this kind takes in function logic the form of a growing sequence of branching structures. However, the process itself is not a branching one because (as was seen in section 7) the order of the crucial steps, all of which are universal instantiations, is irrelevant. All that matters is the introduction of certain constant terms forming a characteristic structure. Because of the independence of quantifiers and connectives discussed in section 4 above, the new constants may be introduced into any formal branch. Since the order of their introductions does not matter, there is a minimum number of instantiations that have to be made in a successful disproof. This number is determined by the formula to be disproved. They match the introductions of new constant terms in successful computation, independently of whether the computation is deterministic or indeterministic.

At any given stage of disproof (model construction) a finite number of constant terms is present. From them one can form a finite number of equations which allow a finite substitution moves. Both numbers are polynomial functions of n.

It follows that in each attempted disproof there is a minimum number of rule applications before it can succeed.

Now in each branch there must occur at least one such substitution move that has not been made already in the initial formula to be disproved. For a formula of the form

(11.3) $\sim(t = t)$

must be introduced to above the branch. And (11.3) could not have been there before a branch started; the branch would then have been closed already.

Hence, the minimum number of rule applications must be at least as large as the number of branches in a successful disproof computation. Then it follows from the Branching Theorem that this minimum length (number of rule applications) must be an exponential function of the length of branches themselves. And this is what P≠NP says.

Essentially the same line of thought can be carried out in terms of more conventional logic instead of function logic. Such an argument is in fact obtainable simply by combining the results of Hintikka 2011(a) and (b).

NOTES

1. See, e.g., Turing, "Computable Numbers"; Church, "Entscheidungsproblem"; Rogers, *Theory of Recursive Functions*; Davis, *Computability and Unsolvability*; Phillips, "Recursion Theory."
2. Hintikka, "Logic as a Theory of Computability"; Hintikka, "Skolem Functions in Proof Theory."
3. Hintikka and Hintikka, *Investigating Wittgenstein*.
4. See, e.g., Hintikka and Sandu, "Game-Theoretical Semantics."
5. Hintikka, "Logic as a Theory of Computability."
6. Hintikka, "Axiomatic Set Theory *in Memoriam*."
7. Hintikka, "No Scope for Scope."
8. See, e.g., Chomsky, *Government and Binding*.
9. See Tarski, *Logic, Semantics, Metamathematics*.
10. Cf. here Hintikka, "Mathematical Logic."
11. See here, e.g., Meinke and Tucker, "Universal Algebra"; Deneke and Wisman, *Universal Algebra*.
12. For the notion of modal set, consult, e.g., Chiswell and Hodges, *Mathematical Logic*.
13. Ibid.
14. As seen in Hintikka, "Skolem Functions in Proof Theory."
15. For universal algebra, see, e.g., Meinke et al. "Universal Algebra"; Denecke and Wisman, ***Universal Algebra***.
16. Cook, "The P versus NP Problem."

BIBLIOGRAPHY

Chiswell, Ian, and Wilfrid Hodges. *Mathematical Logic*. New York: Oxford University Press, 2007.

Chomsky, Noam. *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht: Foris, 1981.

Church, Alonzo. "A Note on the Entscheidungsproblem." *Journal of Symbolic Logic* 1 (1936): 40–41.

Cook, Stephen. "The P versus NP Problem." In *The Millenium Prize Problems*, edited by James Carlson, A. Jaffee, and A. Wiks. 87–104. Providence, RI: American Mathematical Society, 2006.

Copeland, B. Jack, ed. *The Essential Turing*. Oxford: Clarendon Press, 2004.

Davis, M. *Computability and Unsolvability*. New York: McGraw-Hill, 1958.

Denecke, Klaus, and Shelly L. Wisman. *Universal Algebra and Co-algebras*. New Jersey and Singapore: World Scientific, 2009.

Hintikka, Jaakko. *Distributive Normal Forms in The Calculus of Predicates*. Helsinki: Societas Philosophica Fennica, 1953.

———. "No Scope for Scope." In *Linguistics and Philosophy* 20 (1997): 515–49.

———. "Logic as a Theory of Computability." *APA Newsletter on Philosophy and Computers* 11, no. 1 (2011a): 2–5.

———. "On Skolem Functions in Proof Theory." In *Logic without Frontiers. Festschrifft for Walter Alexander Carnielli*, ed. Jean-Yves Beziau and Marcelo Esteban Coniglio. London: College Publications, 2011b.

———. "Which Mathematical Logic Is the Logic of Mathematics?" *Logica Universalis* 6 (2012): 459–75.

———. "Axiomatic Set Theory *in Memoriam*." Forthcoming (b).

Hintikka, Jaakko, and Gabriel Sandu. "Game-Theoretical Semantics." In *Handbook of Logic and Language*, ed. J. van Benthe and A. ter Meulen. 361–410. Amsterdam: Elsevier, 1996.

Hintikka, Merrill, and Jaakko Hintikka. *Investigating Wittgenstein*. Oxford: Basil Blackwell, 1986.

Mann, Allen, Gabriel Sandu, and Merlije Sevenster. *Independence-Friendly Logic: A Game-Theoretical Approach*. Cambridge: Cambridge University Press, 2011.

Meinke, K., and J. V. Tucker. "Universal Algebra." In *Handbook of Logic and Computer Science*, vol. 1, ed. S. Abramsky, Dov M. Gabbay and T. S. E. Maibaum. 189–411. Oxford: Clarendon Press, 1992.

Phillips, I. C. C. "Recursion Theory." In *Handbook of Logic and Computer Science*, vol. 1, ed. S. Abramsky, Dov Gabbay and T. S. E. Maibaum, 79–187. Oxford: Clarendon Press, 1972.

Rogers, Jr., Hartley. *Theory of Recursive Functions and Computability*. New York: McGraw-Hill, 1967.

Tarski, Alfred. *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press, 1956.

Turing, Alan. "On Computable Numbers, with an Application to the Entschendungsproblem." *Proceedings of the London Mathematical Society* 42 (1936): 230–65.

Measuring a Distance: Humans, Cyborgs, Robots

Keith W. Miller
UNIVERSITY OF MISSOURI–ST. LOUIS

David Larson
UNIVERSITY OF ILLINOIS–SPRINGFIELD

BASIC CONCEPTS

Popular notions (as reflected in Wikipedia[1]) place cyborgs directly "between" humans and robots.

> **Humans** (*Homo sapiens*) are primates of the family Hominidae, and the only extant species of the genus Homo. Humans are characterized by having a large brain relative to body size, with a particularly well-developed neocortex, prefrontal cortex, and temporal lobes, making them capable of abstract reasoning, language, introspection, problem solving, and culture through social learning.
>
> A **cyborg**, short for "cybernetic organism," is a being with both organic and cybernetic parts. See, for example, biomaterials and bioelectronics. The term cyborg is often applied to an organism that has enhanced abilities due to technology, though this perhaps oversimplifies the necessity of feedback for regulating the subsystem. The more strict definition of cyborg is almost always considered as increasing or enhancing normal capabilities.
>
> A **robot** is a mechanical or virtual artificial agent, usually an electro-mechanical machine that is guided by a computer program or electronic circuitry. Robots can be autonomous, semi-autonomous, or remotely controlled and range from humanoids such as ASIMO and TOPIO to nano robots, "swarm" robots, and industrial robots.

Another term we will use in this paper is "**artifact**." We define an artifact as something (which could be physical, such as a robot, or logical, such as a computer program) that people create artificially (not, for example, growing it from a seed).

We contend that cyborgs, and their relationships with humans and robots, are worthy of philosophical and practical investigation. Additionally, we will discuss the need for, and problems with, trying to measure a "distance" of a cyborg from a 100 percent human and from a 100 percent robot.

We think this discussion is important because of rapid advances in technology that can be used in conjunction with humans to improve their performance and often their quality of life. It is our contention that as technologies (such as artificial limbs, pacemakers, and other devices) are added to a human, the human then becomes a cyborg. The question then becomes, how much of a cyborg is a particular person with particular artificial replacements and enhancements? We can also approach this issue from the other direction. It seems consistent that a robot can be transformed into a cyborg by adding biological parts. Again, how can we quantifiably relate the resulting entity to a human and to a robot?

Being able to somehow measure a distance between a human, a cyborg, and a robot brings up philosophical issues as well as practical issues. A central philosophical issue is how we define personhood; as an entity moves from 100 percent human by replacing or adding mechanical parts, is there a point at which that entity is no longer a person? If there is not such a point, then does that automatically mean that sufficiently sophisticated machines will inevitably be classified as persons? Practical issues include implications in sports, health care, health insurance, life insurance, retirement policies, lawsuits, discrimination, and in software design and implementation for cybernetic devices.

MEASURING THE DISTANCE

If you look at the problem of measuring humans – cyborgs – robots from a purely physical view, it seems logical that there exists a continuum from 100 percent human to 100 percent robot with cyborgs being the transition from one to the other. At the moment, cyborgs typically start out as humans and mechanical parts are added. However, there is nothing in our notion of cyborgs that would theoretically prevent moving in the opposite direction: adding biological parts to robots.

No matter how a cyborg comes into being, we would like to be able to talk about a "distance" from a given cyborg to the ends of the continuum: 100 percent human and 100 percent robot. It's easy to draw a picture that depicts the basic idea (see Figure 1), but it's not so easy to define precisely what the distance in this picture means. We contend that we need a metric to mark the distance between the extremes, a metric that is exactly correct at both extremes but gives appropriate measures of cyborgs, entities that are part human and part machine. Some may object that it has not been established that there should be a linear relationship that can be measured; for example, there may be several different observables that should be taken into account, so that the "distance" would be based on a vector rather than a scalar. We will consider this possibility later in this paper, but for now we will assume the scalar continuum and explore what progress we can make on establishing this measure.

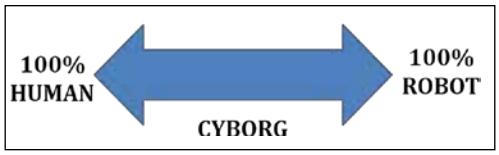

Figure 1. Cyborgs are mixtures of biological human parts and mechanical parts.

In the following discussion we will examine problems we have encountered in trying to measure the distance from humans to robots. We find the movement from non-artifact to artifact to be particularly interesting and believe a marvelous example of this movement occurs when a human being adds mechanical or electro-mechanical parts to become a cyborg. We will look carefully at that movement, and suggest different ways that we might measure the distance from 100 percent human to 100 percent robot. If we can define an

effective measure for this movement, that measure will have both theoretical and practical significance.

MORE ON CYBORGS

Cyborgs are a popular theme in fiction. The *6 Million Dollar Man* and the *Bionic Woman*, seven of nine from *Star Trek*, and Detective Spooner in *I Robot* are well-known fictional characters that are, by our definition, cyborgs.

There are non-fictional cyborgs as well. Scientists Steve Mann[2] and Kevin Warwick,[3] both leading researchers in cybernetics, are also cyborgs because of the mechanical parts they have inside and outside themselves. These parts include RFID chips under the skin and glasses that augment visual reality, an idea now taken up by Google. Another well-known cyborg is Oscar Pistorius, a runner from South Africa. While Steve Mann and Kevin Warwick are cyborgs by choice, Pistorius became a cyborg because he needed a replacement for both his legs below the knees.[4]

Running on his spring steel artificial legs, Pistorius was a successful runner in the Special Olympics. But Oscar wanted to race in the Beijing Olympics. Some of his potential competitors objected, contending that his artificial legs were more efficient than human legs, and that therefore Pistorius would have an unfair advantage. Oscar sued and the courts overruled the Olympic Committee. Pistorius was allowed to compete for a spot on the South Africa Olympic team.

Pistorius, nicknamed "the Blade Runner," is a classic case of a cyborg: part human, and part mechanical. Pistorius's legs are not automated but they are artificial. They are also "attached" rather than being internal. However, especially when he is competing, Pistorius is not 100 percent carbon-based human, and he is not 100 percent mechanical. Pistorius's case brings up an interesting issue: Does the cybernetic part enhance the person and provide them with more than "normal" capabilities, or does it just make them "normal"?

There are many cyborgs among us today. Artificial parts are becoming increasingly common, and those parts are increasingly sophisticated. They deliver clear advantages to people with impairments and missing limbs. In some cases, these devices are being used to enhance, rather than replace, human functions.

In a recent paper in IEEE Technology and Society,[5] Roger Clarke writes about cyborgs and possible legal issues under the title "Cyborg Rights." Clarke reviews several aspects of how cyborgs are defined and categorized. The kinds of artifacts used to make a cyborg can be distinguished by their intent: Are they prosthetic, meant to replace missing or diminished functionality; or are they orthotic, meant to enhance normal functioning? Clarke also separates the artifacts by their relationship to the body of the cyborg. Any artifact that is under the skin he calls *endo*. A cardiac pace maker and a cochlear implant are endo. An artifact that is attached to the body but not inside the body is labeled *exo*. Oscar Pistorius's legs are exo. The third category, *external*, includes devices that are not inside and not attached to the body, but are still integrated with the human body.[6] Eyeglasses, canes, and scuba gear are external.

Clark explores six different kinds of cyborgs using the distinction we've already drawn between prosthesis and orthotic, and the distinction among endo, exo, and external. An example of an endo prosthesis is an artificial hip. An example of an orthotic endo device would be a metal plate attached to a bone to make it stronger than a normal human bone. An example of an exo prosthesis is an artificial leg, such as the legs Oscar Pistorius uses. If an artificial leg is somehow mechanically superior to a human leg for some particular function, then it is orthotic instead of a prosthesis, but it is still exo. A pair of eyeglass and a cane are both external prosthesis devices. Devices used to enhance human vision, such as microscopes or telescopes, are external, orthotic devices.

If you agree with Clarke that eyeglasses, contact lenses, and scuba gear all make you a cyborg, then many of us are best classified as cyborgs. This inclusive view of cyborgs is attractive because it emphasizes a broad range of ways in which we can augment ourselves artificially. But since the word "cyborg" came from "cybernetics," a term popularized by Norbert Weiner with respect to information flows in systems, some scholars may be more comfortable with restricting cyborgs to devices that are electromechanical.

Examples of existing electromechanical devices used with humans are a cardiac pacemaker and embedded RFID chips (common in pets but also used in humans); both of these are endo devices. A robotic hand can be an exo prosthesis or an exo orthotic, depending on how much functionality it delivers. An artificial lung not inside a patient is an external prosthesis device.

In trying to establish a metric for measuring the distance between humans and robots, we found it difficult to include external devices in our initial analyses. External devices can be too temporary, too loosely integrated into the person, to be considered (at least by some) to be well integrated with the original person. When considering the measurement question, external devices led us to some difficulties. For example, including external devices in the definition of "cyborg" made it difficult to precisely distinguish a person using a tool from a cyborg that has integrated a mechanical external part. For now, we will focus on cyborgs made with endo and exo devices; devices attached to the body and embedded under the skin have a clear distinction from tools "at hand." However, others may wish to be more inclusive than we have been and explore alternative metrics.

CANDIDATES FOR MEASURING THE DISTANCE

Now that we have clarified the scope of what we will consider as qualifying devices to move someone along the cyborg continuum, we will focus on our original quest—determining a measure of cyborg-ness. All of the measures will give a reading of 0 percent for 100 percent humans, and give 100 percent for a 100 percent robot.

CANDIDATE 1: BY WEIGHT

The first measure is simple, relatively easy to measure, and sadly counter-intuitive. In this strategy, we divide the weight of a cyborg's mechanical parts by the total weight of the cyborg, and multiply that ratio by 100. If you have a hip replacement, that moves you to the right from 0 percent. A heart valve replacement moves you further. If someday

you add an artificial memory implant or artificial lungs, those changes would move you further still on the continuum.

As with all the measures we will consider, there are some good and bad aspects to the weight metric. For this weight-based measure, one advantage is that you can determine this measure precisely in an objective, straightforward way. A significant disadvantage is that the weight of a part is intuitively not indicative of the significance of the replacement. A brain weighs about 1.3 kilograms (about three pounds). A leg weighs about seven kilograms (about fifteen pounds). Realistically most people do not think a leg is five times as important as the brain, but according to the weight measure, replacing a leg with a mechanical device would place you more towards the robot side of the continuum than replacing your brain with a mechanical device.

CANDIDATE 2: BY INFORMATION FLOW

The next measure considered is an information-centric measure. For this measure, the information flow into and out of each replacement part in the cyborg is determined. We also measure all the information flow in the whole body. The ratio gives us our position on the continuum. One good aspect of this measure is that, in theory, it could be measured and information flow could potentially relate to significance. (For example, we might be able to approximate the number of bits necessary to contain the same information that travels through the nervous system in and out of the artificial part and/or the biological part replaced.) Also, this measure is consistent with Floridi's emphasis on information as particularly significant in understanding ethical significance.[7] An unfortunate aspect of an information flow metric is that it may be difficult to measure. Although we may be more capable of this measure sometime in the future, it is not currently practical to precisely measure flow for all parts. One complication is that the nervous system is not the only way the body communicates; for example, hormones are part of a chemical-based communication that is not restricted to the nervous system. This would have to be taken into account for an accurate measurement. Another complication is that we would have to take into account the possibility of redundant and irrelevant information—Should such information be included or excluded from our measure? (Thanks to editor Peter Boltuc for pointing out this potential complication.)

A problem with the "by weight" measure is that it over-emphasized the importance of mass. The "by information flow" may have a related problem: it perhaps over-emphasizes the significance of information processing to humans and de-emphasizes other aspects, such as mobility and disease prevention. If the measure was restricted to information flow in and out of the brain, that might make it more practical to measure but less sensitive to important aspects of humanness.

CANDIDATE 3: BY FUNCTIONALITY

The third measurement option adds up the functionality contributed by the artificial parts and divides that by the total functionality of the cyborg. This avoids the counter-intuition of the weight measure and allows us to include more considerations than information flow. But, as stated, this is a vague measurement and measuring precisely the "amount of functionality" is difficult. We have by no means solved these problems, but looking at attempts to measure medical outcomes of injuries and therapies may offer us directions for future research in measuring "cyborg-ness."

For many reasons, including insurance payments and scholarly studies about effective treatments, medical professionals seek to quantify patients' quality of life. This has led to numerous attempts to assign numerical values to a patient's physical and mental well-being.[8]

One way to attempt a functionality measurement is to adopt or adapt one of the existing systems that helps assign a number to the impairments a patient exhibits, or to the improved condition of a patient. The adaptation would have to isolate and then measure the effect of the cyborg's improvement due to the endo or exo artifact that was added. These measures attempt to measure both time (increased life span) and quality of life.

A positive aspect of using (or adapting) an existing quality of life measure is there is an extensive body of literature and years of active practice in using these measures. There has been considerable effort to make these evaluations repeatable, consistent, and objective. However, there is not as yet a consensus on any one particular measure, and the objectivity of these measures is an ongoing object of research. A significant problem for our purposes is that the measures now in use necessarily use statistical measures of large groups of patients, and the effect of an integrated artifact may vary significantly between individual cyborgs.

The use of these medical function-based measures for quality of life seems more clearly appropriate for prosthetic artifacts than for orthotic enhancements. For prosthetic devices, a certain function that was human becomes mechanical, and that gives rise to a movement to the right in our diagram, a distance proportional to the percentage of functionality. However, an orthotic enhancement introduces new functionality and perhaps longer life. Seemingly, this is significant for quality of life type measures. However, how do we reflect these improvements in a measure for cyborg-ness? Reflecting these improvements could, for example, push us beyond 100 percent robot, which intuitively makes no sense. Perhaps a way out of this problem is to restrict our measurement to the percentage of functionality of the replaced or enhanced part based on a 100 percent human. But that restriction clearly loses some of the explanatory power of functionality-based measuring.

The intuitive appeal of measuring functionality has convinced us that this area requires more study. However, the complications involved will require the involvement of medical professionals in order to make a more intelligent suggestion about how these measures might be adapted to measuring the distance between humans, cyborgs, and robots.

CONSIDERING CHALLENGES TO THE PROPOSED MEASURES

We are not overjoyed with any of the alternatives listed above; no measure seems ideal. Perhaps some measures can be combined, but that adds unwanted complexity. If the measure is too complex, people will not readily understand

it, and fewer people and institutions are likely use it. We hope, therefore, to be able to use a single measure if at all possible.

A challenge for any cyborg measure is trying to include sensitivity towards how an artifact is used by a human, not just what the artifact and its capabilities are. As an example of this challenge, consider what we call the "surrogate problem." The idea of artificial entities under the control of human operators was explored by a recent movie called *Surrogates*, and in a different way by the more commercially successful (but less well explained scientifically) movie called *Avatar*. The surrogate problem for our cyborg metric is that the degree to which external devices are used by the humans as a substitute for life without the surrogate should be a significant factor in our measure. A person who spends almost every waking hour living "through the surrogate" seems clearly more of a cyborg than a person who uses a very similar device, but uses the device sparingly, a few minutes a day. The purpose and duration of these surrogate sessions should help determine our measure of cyborg-ness, not just the artifacts themselves. Our measures above, which exclude external devices, do not wrestle with this issue, but it clearly is an issue worthy of further study.

The surrogate problem is related to another problem with external devices, which we call the "puppet problem." We referred previously to the problem of distinguishing between a tool and an external integrated artifact. One prominent existing example of what might be classified as an orthotic external device is a Predator military drone. These drones are already ethically and legally controversial and thinking of the operators and drones collectively as a cyborg adds another layer of complexity to the ethical questions. Human/machine collaborations (like a physical puppet) become more complex and more significant when the "puppet" has onboard intelligence. Predator drones started out as electronic puppets, controlled in a way that is similar to video games. But as the drones become increasingly sophisticated, they are gaining more and more internal control, and plans have been developed to make them independent from direct human control for longer and longer times.[9] A challenge for adapting a metric to external puppet devices is how to measure the sophistication of an artifact, and the degree of control the human has over the mechanical artifact. The degree of sophistication and independence of an artifact in a "puppet cyborg" seems like a significant factor in understanding the cyborg as a whole, but artifact sophistication and independence are difficult to quantify.

These questions of measuring puppets also can be applied to endo and exo artifacts. Should not our metric take into account the intelligence and perhaps the "autonomy" of the artifacts that are part of the cyborg? If so, how exactly should this be measured? If not, then aren't we missing a potentially significant aspect of the cyborg? For example, if a cyborg includes an endo computer that can override the brain's signals in case of an emergency, or in the case of a detected malfunction in the brain, then that automatic control appears to be an important movement towards the robot side of the continuum that should be included in our measure. None of the measures proposed above is sensitive to this kind of distinction.

A separate challenge has to do with the history of how a cyborg is constructed. We explicitly draw our arrow between humans and robots as a double-headed arrow, but typically the literature discusses a cyborg as a human plus mechanical parts. That is certainly an interesting direction, and it is the direction we are going with many experiments in cyborgs today. But it isn't the only direction possible. We can also make cyborgs by going in the other direction, placing human parts into robots and making a cyborg that way (see Figure 2). This has interesting ramifications. Our measures above could work for these "left-handed cyborgs," but people may feel quite differently about a robot with added biological parts than a human with added mechanical parts, even if the measurements of the two resulting entities are identical. On the basis of fairness, we suspect that the measure should not be directly sensitive to the history of the cyborg's formation, although a measure could be sensitive to different functionalities that resulted from different formations of a cyborg.

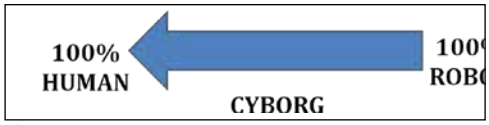

Figure 2. A cyborg could be built by adding biological parts to a robot.

FURTHER IMPLICATIONS

We have only started to touch on several important questions about measuring cyborgs. Once we establish a measure (or measures), that is only the beginning of the philosophical work. With a measure established, you then have to face difficult questions. For example, is there a particular place on your scale where you start to restrict how a cyborg is treated by people or the law? Is there some point at which this someone should not be called "human" anymore? If so, what is that number, exactly? If not, then when all the biological parts are replaced, is the resulting entity (which on our account is now a robot) still a person?

Should voting rights change if you become too close to a robot? If you live hundreds of years as a cyborg, and you vote every year, then you will get many more votes than a human limited to about a century. Is that acceptable, especially for 100 percent humans? How will health care financing change when cyborg replacement parts become a huge part of health spending? Will people who are NOT cyborgs become a distinct minority? If so, humans without artificial enhancements may find themselves living shorter lives than cyborgs, and cyborgs (with increased life span and enhanced capabilities) may routinely surpass non-cyborg humans in business and government; how will society adapt to these kinds of changes? Will non-cyborg humans insist on a legal status separate from cyborgs? What are the ethical and political arguments for and against making this distinction in laws and regulations?

Many scholars are becoming interested in how we are being transformed from a society of humans to a society that combines humans and other intelligent entities. We think that we can make progress in exploring these issues by concentrating on cyborgs. First, unlike other entities being discussed (like robots who could readily pass for humans

in a physical encounter), cyborgs are already among us in large numbers. Second, in the foreseeable future it is likely that many of us will be moving towards robots in the human-robot cyborg continuum. Increasing numbers of people will have an immediate, personal stake in these questions about cyborgs.

CONCLUSIONS

In this short article we have asked more questions than we've answered. But there are several ideas that seem clear:

1. Cyborgs are among us, and the number of cyborgs is likely to increase. The sophistication of the mechanical parts will continue to increase.

2. Software used to control the mechanical parts will become more sophisticated and complex.

3. The idea of a continuum from 100 percent human to 100 percent robots can be a useful notion in our philosophical and ethical analyses, even without settling on a particular metric.

4. Practical problems in making policies, laws, and regulations will not be well served by a theoretical, under-specified continuum. Policy must be spelled out in a way that is unambiguous and precise, so for any given law or regulation that establishes rules based on cyborg-ness, a particular measure (or perhaps measures) will have to be chosen.

5. Despite the many challenges to determining a fair and accurate measure, we are convinced that work should continue on establishing a metric to measure cyborg-ness.

We plan to ask exactly these kinds of questions in our future work.

APPENDIX A: HUMAN CLONING

If someday we have human clones, they will seriously complicate our view of a cyborg continuum between completely biological and completely artificial. Most people today would not think that human clones are identical to more traditional biological humans, but they will not necessarily have any mechanical parts. So how then do human clones relate to our cyborg measure? Clones are artificial in a significant way, but they are not mechanical at all. The classic science fiction movie *Blade Runner* wrestles with ethical and legal issues that might arise if human clones become common. Perhaps what we need to do in the case of human cloning is to isolate the case of human cloning with its own continuum.

Consider a new continuum where all the entities on the continuum are biological (not mechanical), and where the two extremes are no cloning on the left and 100 percent cloned on the right (see Figure 3). As far as we know there is not a consensus term for humans who have replaced some original parts with cloned parts, so we have marked the middle of this continuum with the phrase "somewhat cloned."

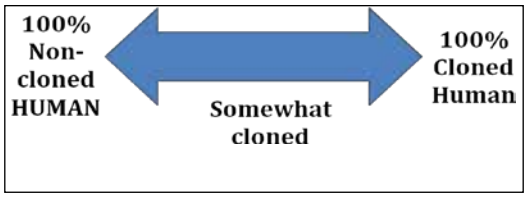

Figure 3. Cloning requires a new scheme for measuring the distance between biological humans and cloned humans.

NOTES

1. See Wikipedia, "Cyborg"; Wikipedia, "Human"; and Wikipedia, "Robot."
2. Monaco. "Future of Wearable Computing."
3. Warwick, http://www.kevinwarwick.com/.
4. As we write this article, the controversies about Pistorius have intensified because of his involvement with a fatal shooting. See Pererira, "Ex-Lead Investigator in Oscar Pistorius Murder Case Convinced He Intentionally Killed Girlfriend."
5. Clark, "Cyborg Rights."
6. The exact meaning of "integrated" is controversial, and beyond the scope of this article. Clarke's article discusses this in some detail (see pages 10 and 11). If "integrated" is interpreted in a way that includes more devices that people use, then more people are properly called cyborgs; if "integrated" is interpreted in a way that excludes devices unless, for example, they are attached permanently to the body, then fewer people should be called cyborgs. In this article we somewhat arbitrarily exclude some devices that are used, but not intimately attached to the body; however, a quite similar analysis could be done with a more inclusive definition of "integrated." Notice that we have not considered, as others have, the idea of drugs as artificial enhancements that should be included when considering cyborgs. See Clynes and Kline, "Cyborgs and Space."
7. Floridi, "Information Ethics."
8. Torrance and Feeny, "Utilities and Quality-Adjusted Life Years"; and Horne and Neil, "Quality of Life in Patients with Prosthetic Legs."
9. Keller, "Smart Drones."

BIBLIOGRAPHY

Clarke, Roger. "Cyborg Rights." *IEEE Technology and Society* 30, no. 3 (2011): 49–57.

Clynes, Manfred E., and Nathan S. Kline. "Cyborgs and Space." *Astronautics* 14, no. 9 (1960): 26–27.

Floridi, Luciano. "Information Ethics: On the Philosophical Foundation of Computer Ethics." *Ethics and Information Ttechnology* 1, no. 1 (1999): 33–52.

Horne, Carolyn E., and Janice A. Neil. "Quality of Life in Patients with Prosthetic Legs: A Comparison Study." *Journal of Prosthetics and Orthotics* 21, no. 3 (2009): 154–59.

Keller, Bill. "Smart Drones." *New York Times Sunday Review*, March 16, 2013, http://www.nytimes.com/2013/03/17/opinion/sunday/keller-smart-drones.html?pagewanted=all&_r=0, accessed July 23, 2013.

Monaco, Ania. "The Future of Wearable Computing." IEEE Institute, January 21, 2013, http://theinstitute.ieee.org/technology-focus/technology-topic/the-future-of-wearable-computing, accessed July 12, 2013.

Pereira, Jen. "Ex-Lead Investigator in Oscar Pistorius Murder Case Convinced He Intentionally Killed Girlfriend." *ABC News*, May 6, 2013, http://abcnews.go.com/International/lead-investigator-oscar-pistorius-murder-case-convinced-intentionally/story?id=19080527#.UYfD77Xvh8E, accessed 16 July 2013.

Torrance, George W., and David Feeny. "Utilities and Quality-Adjusted Life Years." *International Journal of Technology Assessment in Health Care* 5, no. 4 (1989): 559–75.

Warwick, Kevin. "The University of Reading." http://www.kevinwarwick.com/, accessed July 12, 2013.

Wikipedia. "Cyborg." http://en.wikipedia.org/wiki/Cyborg, accessed March 26, 2013.

———. "Human." http://en.wikipedia.org/wiki/Human, accessed March 26, 2013.

———. "Robot." http://en.wikipedia.org/wiki/Robot, accessed March 26, 2013.

The Ethics of Creating Artificial Consciousness

John Basl
NORTHEASTERN UNIVERSITY

1. INTRODUCTION

The purpose of this essay is to raise the prospect that engaging in artificial consciousness research, research that aims to create artifactual entities with conscious states of certain kinds, might be unethical on grounds that it wrongs or will very likely wrong the subjects of such research. I say *might be unethical* because, in the end, it will depend on how those entities are created and how they are likely to be treated. This essay is meant to be a starting point in thinking about the ethics of artificial consciousness research, not, by any means, the final word on such matters.

While the ethics of the creation and proliferation of artificial intelligences and artificial consciousnesses has often been explored both in academic settings and in popular media and literature, those discussions tend to focus on the consequences for humans or, at most, the potential rights of machines that are very much like us.[1] However, the subjects of artificial consciousness research, at least those subjects that end up being conscious in particular ways, are research subjects in the way that sentient non-human animals or human subjects are research subjects and so should be afforded appropriate protections. Therefore, it is important to ask not only whether artificial consciousnesses that are integrated into our society should be afforded moral and legal protections and whether they are a risk to our safety or existence, but whether the predecessors to such consciousnesses are wronged in their creation or in the research involving them.

In section 2, I discuss what it means for a being to have moral status and make the case that artificial consciousnesses of various kinds will have moral status if they come to exist. I then take up the issue of whether it is thereby wrong to create such entities (section 3). It might seem obvious that the answer is "no," or at least it is no more impermissible than the creation and use of non-human research subjects. However, I argue that there should be a presumption against the creation of artificial consciousnesses.

2. MORAL STATUS AND ARTIFICIAL CONSCIOUSNESS

In order to determine whether it is possible to wrong artificial consciousnesses by creating them or conducting research on them, we must first determine whether such entities have moral status and what the nature of that status is.

2.1 WHAT IS MORAL STATUS?

The term "moral status" is used in various ways in the ethics and applied ethics literature. Other terms, such as "inherent worth," "inherent value," "moral considerability," etc., are sometimes used as synonyms and sometimes to pick out species of moral status.[2] In the broadest sense of the term, to have moral status is just to have any kind of moral significance; that is, having moral status means that in at least some contexts moral agents must be responsive to or regard the thing that has moral status.

It would be very easy to argue that artificial consciousnesses have moral status in the broad sense just described. After all, even a rock, if owned by someone, or part of a piece of art, for example, has moral status in this sense. Instead, I will employ the term "moral patient" to pick out a particular form of moral status. The definition of "moral patient" as used in this paper is:

> **Moral Patient**$_{df}$: X is a moral patient iff agent's like us are required to take X's interests into account in our moral deliberations for X's sake when X's interests are at stake.

This definition has the following features:

1. A being is a moral patient only if it has *interests* that are to be taken into account in moral deliberations.

2. A being's being a moral patient entitles it have its interests taken into account in moral deliberations for *its own sake*.

3. Moral patiency is a property had by an entity *relative to agents like us*.

Each of these features will be discussed in detail below, but first it is important to discuss the relationship between moral patiency and normative theory. Some view the question of whether a being is a moral patient as dependent on which normative theory is true.[3] That is, in order to determine which beings are patients, we must first figure out whether we should be, for example, Utilitarians or Kantians, Virtue Theorists or Contractualists. If this thesis, call it the Dependency Thesis, about the relationship between moral status and normative theories is correct, we can't answer the question of whether artificial consciousnesses are moral patients without first answering the question of which normative theory is correct.

There are important relationships between normative theory and moral status. For one thing, which normative theory is true explains the nature or source of the moral status of whichever beings have it. If contractualism is true, for example, a being's moral status is grounded in or finds its source in the consent of rational contractors; if utilitarianism is true, a being's moral status is grounded in the fact that it's being benefitted or harmed contributes to or detracts from the value of a state of affairs. Furthermore, how, in particular, moral patients are to be treated is a function of which normative theory is ultimately correct. Utilitarianism licenses the killing of moral patients more easily than a Kantian ethic, for example. For this reason, the strength of the presumption against creating artificial consciousnesses defended below will depend on which normative theory is true. However, the Dependency Thesis concerns relationship between normative theory and moral patiency with respect to *which beings are moral patients*.[4]

Fortunately, the version of the Dependency Thesis that precludes us from determining whether artificial consciousnesses are moral patients independently of

determining which normative theory is true is false. One point in favor of thinking that it is false is that we know that all adult humans of sound mind are moral patients, and yet we aren't sure which normative theory is true, or, at least, whether all adult humans of sound mind are moral patients is far less controversial than which normative theory is true.

One might argue that the obviousness of our patiency just serves as a condition of adequacy on normative theories and that's why we know we are patients even if we haven't settled which normative theory is true. However, it also suggests the possibility that we can make a similar case for the moral status of other beings. That is, even if some metaphysical, ontological, or supervenience version of the Dependency Thesis is true, we may have ways of specifying which things are moral patients independently of determining which normative theory is true. All that really matters for the purposes of arguing that artificial consciousnesses are or can be moral patients is that the dependency relationship between patiency and normative theory isn't epistemic (i.e., so long as we can come to know that some being is or isn't a moral patient without determining which normative theory is true).

There is good reason to think we can come to know who or is a moral patient independently. Debates about which entities have moral status and about the degree to which entities of various kinds matter happen, as it were, internal to normative theories. Utilitarians, for example, have argued about whether non-human animals and human infants are moral patients on par with us.[5] There are some Kantians that argue that many non-human animals should be accorded many rights in the same way that we ought.[6] So long as the intra-normative debates are coherent we can be sure, at least, that normative theories aren't fully determinate of which beings have moral status.

Furthermore, the kinds of arguments made that this or that entity is a moral patient do not typically appeal to which normative theory is true.[7] Consider, for example, a standard argument from marginal cases that non-human animals have moral status. Such arguments take for granted that so-called "marginal cases," such as infants and the severely mentally handicapped, have moral status. Then an argument is made that there is no morally relevant difference between marginal cases and certain non-human animals, for example, chimps. From this it is concluded that chimps are moral patients in the same way that we are. This argument doesn't make explicit mention of normative theory, nor do the arguments typically given for the premise that there is no morally relevant difference between chimps and marginal cases.

I'm not here endorsing any particular argument from marginal cases or assessing its merits. The point is that the kinds of arguments that a Utilitarian might use to convince another Utilitarian that chimps matter are the same kinds of reasons that should convince a Contractualist or Kantian to accept that chimps are moral patients. Similarly, if Kantians could make a case that, for example, only the interests of very cognitively advanced beings are relevant to moral deliberations, that advanced cognitive capacities are a morally relevant properties, they won't do so by appealing to the structure of Kantian normative theory, but to reasons that a Utilitarian could accept; at least they will do so if they hope to convince other Kantians that don't share their view about the relevance of advanced cognitive capacities.[8]

The above considerations provide an abbreviated, but I hope sufficient, case for the idea that we can identify moral patients without first discovering which normative theory is true.

2.1.1 INTERESTS

According to the definition of moral patiency, if a being is a moral patient we must take that beings interests into account for the sake of that being. To say that a being has interests is to say that it has a welfare, that it can be benefitted or harmed.[9] Whether a being is potentially a moral patient depends, therefore, on whether it has a welfare, and that depends on which theory of welfare is true.

There are various families of views about welfare and some are more stringent about the features a being must have to have a welfare.[10] I don't intend here to settle the issue of which theory of welfare is true. Instead, below I will focus on a type of artificial consciousness that will have a welfare independently of which of many plausible theories of welfare is true.

A being's welfare can be significant in moral deliberations for a variety of reasons. For example, if I hire a dog walker, they have an obligation to me to take my dog's interests into account. However, they also, I contend, have an obligation to take my dog's welfare into account for her sake; even if I didn't own my dog, even if no one does, it would be wrong for the dog walker to kick my dog for no reason.[11]

Some being's welfare may only matter derivatively, but a moral patient's welfare matters for its own sake.[12] The interests of a patient figure into our deliberations independently of their relationship to the welfare of others.[13]

2.1.2 THE AGENT RELATIVITY OF MORAL PATIENCY

It might seem odd, even contradictory, to claim that a moral patient's welfare matters in moral deliberations for their own sake while at the same time also relativizing moral patiency to a set of agents like us. However, rather than being contradictory this reflects the fact that agents that are radically different from us might exist in an entirely different ethical world, so to speak.

Let's imagine, for example, that there is a type of being that is completely immaterial. Admittedly, I don't know how to understand how such beings interact in any sense, but I do know that whatever such beings do, they cannot have any effect on beings like us and so they are not required to take our welfare into account in whatever moral deliberations they have.

Or, assume that Lewis was right and that all possible worlds really exist in the normal everyday sense of exists.[14] There are worlds very much like ours that are, in principle, causally cut off from us. The moral agents in those possible worlds are under no obligation to take our welfare into account because they can't affect us in any way.

Finally, imagine that rocks have a welfare but that it is impossible for us to come to know about that welfare. In

such a case, while we may make these beings worse off, we are either under no obligation to take their welfare into account, or if we are so required, we are excused for failing to do so because of our ignorance, and so for all practical purposes rocks are not moral patients.[15]

These examples show, at least in principle, that whether a being is a moral patient is agent relative; it is relative to agents sufficiently like us that engage in causal interactions with potential patients and which can come to know or have reasonable beliefs that their actions affect the welfare of potential patients.

2.2 CAN ARTIFICIAL CONSCIOUSNESSES BE MORAL PATIENTS?

There is not a single question of whether artificial consciousnesses could satisfy the conditions of moral patiency. There is a technological version of the question: Will we ever be in a technological position to create artificial consciousnesses that satisfy the conditions of patiency?

The answer to that question depends in part on an answer to a nomological version of the question: Do the laws of our universe make it possible to create consciousness out of something other than the kind of matter of which we are composed and configured in a way that's very similar to consciousnesses we know of?

The technological and nomological questions just raised are interesting and important, especially to those who wish to create artificial consciousnesses. However, as a philosopher, I'm in no position to answer them. I'm going to assume that artificial consciousnesses with a large range of cognitive capacities are creatable and instead focus on the following conceptual question: Is it conceptually possible to create an artificial consciousness that is a moral patient?

I think the answer to this question is clearly "yes." To see why, just imagine that we've managed to create an artificial consciousness and embodied it, certainly a conceptual possibility. This being is, we know, mentally very much like us. It is a moral agent, it has a similar phenomenology, it goes about the world much like we do, etc. What would we owe to this being? I think it is our moral equal and that denying that would make one, to use Singer's term, a *speciesist*. But even if you think that such a being would not be our moral equal, it would certainly be wrong to hit such a thing in the face with a bat, or to cut off its arm because of the effect such actions would have on the welfare of such a being. That is, even if we have some special obligations to the members of our own species and some degree of partiality justified, this kind of artificial consciousness is a moral patient.

The more interesting question isn't whether an artificial consciousness very much like us is a moral patient, but what are the minimal conditions for an artificial consciousness to be a moral patient. After all, it seems plausible that merely being conscious does not make a thing a moral patient. Imagine that we can create an artificial being that has conscious experiences of color but nothing else.[16] Such a being is not a moral patient because it doesn't have a welfare; which color experience it is having, by hypothesis, doesn't make its life go better or worse.

So what are the minimal conditions for an artificial consciousness to have a welfare that we should care about for its own sake? That's a much harder question to answer. Fortunately, we can say something about the ethics of creating artificial consciousnesses without fully answering it by thinking about existing non-human moral patients.

It is, I hope, relatively uncontroversial that at least some non-human animals, mammals and birds in particular, are moral patients. It is at least pertinent to our deliberations about whether to experiment on such animals, whether it is permissible to withhold food from our pets, to encourage them to fight for our pleasure, that these activities will affect the welfare of these beings (and that welfare of such beings is pertinent for their own sake). It is more controversial whether such beings matter because they have the capacity for suffering and enjoyment, whether they have desires that can be satisfied or frustrated, whether they are sufficiently rational, etc. That is, it is controversial in virtue of which properties they have a welfare, but relatively uncontroversial whether their welfare ought figure into our moral deliberation for the sake of those animals whose welfare is at stake.

For the purposes of evaluating the ethical case against the creation of artificial consciousnesses, we can restrict that evaluation to the creation of artificial consciousnesses with capacities similar to non-human animals that we take to be moral patients. If it turns out that the conditions for a being a moral patient are less restrictive, the conclusions below will apply to artificial consciousnesses of that type.

3. THE CASE AGAINST CREATING ARTIFICIAL CONSCIOUSNESS

Just as there is nothing intrinsically wrong with creating biological consciousnesses through traditional means (i.e., breeding animals or having children), there is nothing intrinsically wrong with creating an artificial consciousness, at least not from the perspective of the created being.[17] If scientists were to create an artificial consciousness like that described above, that goes about the world as we do and has a mental life like ours, and those same scientists and society generally were to treat that being in a way that was commensurate with its being a moral patient, that would be morally permissible. However, there are reasons to expect that such a being would not be treated in a way that is commensurate with its being a moral patient and in such a case, we have good reason not to allow the creation of such a consciousness, at least not without adequate protections. The case against the creation of artificial consciousnesses is thus conditional: there is an ethical reason not to create artificial consciousnesses when there is sufficient risk that such beings will be moral patients and when there is also sufficient risk that these patients will be mistreated.

To illustrate the kinds of risks to artificial consciousnesses associated with their creation it is useful to first discuss some cases having to do with traditional organisms. Consider the following case:

> **Neural chimera**: Researchers are attempting to create human-animal neural chimeras by injecting human stem cells into the brains of guppies. In light of some recent developments in stem-cell research

> and in neuroscience, the scientists think that they can significantly alter the cognitive capacities of the resultant guppies by doing so. Their hope is to create guppies with brains much more like ours that they can use in Parkinson's research. Since guppies are relatively cheap to feed and reproduce quickly, they think this would be an excellent solution to the need for better animal models for Parkinson's research.

Let's assume that researchers, after conducting the research described, intend to care for the resulting guppies in the same way that they care for typical guppies. If so, such research is almost certainly unethical. If the scientists are right that there is a significant chance that the resulting guppies will be, mentally, a lot more like us, they would be owed much more than what is typically accorded to guppies in normal research contexts. By creating a moral patient that is much like us, the scientists obligate themselves to treat these subjects commensurately with that moral status, but the research as described fails to do so.

This kind of case seems far-fetched, but scientists are concerned with the creation of such chimeras, and these sorts of ethical worries have been raised by others about this research.[18] And it is not the only sort of research that raises these worries. I've argued elsewhere that testing cognitive enhancements on non-human research subjects has the potential to alter their capacities in ways that increase the risks that they will be mistreated.[19]

The case above serves to illustrate what might make creating artificial consciousness unethical. That's not to say all artificial consciousness research will be unethical. In assessing the ethics of creating artificial consciousness research programs from the perspective of the research subjects (the artificial consciousnesses that might be created) the following questions must be addressed:

1. How probable is it that a given research program will result in the creation of an artificial consciousness that is a moral patient?

2. How probable is it that such patients will fail to be treated appropriately?

These questions are difficult to answer in any precise fashion. It will vary from research program to research program, and it will depend on what safeguards are put in place to protect the interests of the research subjects. Still, there are some considerations that suggest that these probabilities are sufficiently high (or at least should be judged to be so).

3.1 THE PROBABILITY OF CREATING ARTIFICIAL CONSCIOUSNESS

How probable the creation of artificial consciousness is in a given research context is extremely difficult to determine. This is in part because it depends on the nature of consciousness. For example, if consciousness requires neural correlates, and those correlates aren't realizable using the methods or materials in use in some research program, the probability of creating an artificial consciousness is low.

The problem is that the nature of consciousness is a difficult problem and so we can't be sure which research programs are most promising with respect to creating artificial consciousness. In fact, it might be that the question of consciousness goes unsettled until researchers are able to create an artificial consciousness to confirm one or other of various theories.

Given these difficulties, what should we say about the probability of creating an artificial consciousness? Are we stuck thinking there is no way to assign a probability one way or another and so need to concern ourselves with the ethical risk to research subjects? I think not. The reason is that every attempt to create artificial consciousness is taken with the aim of success and because of the ethical risk success carries.

Attempts to create artificial consciousness will not be made at random. Researchers will attempt methods they think promising or that have a better chance of success than alternatives. This doesn't tell us that the probability that an artificial consciousness of the sort that would be a moral patient will be created is high, but I think that it provides us a reason to assume that it is high if there are ethical risks associated with success. That is, since artificial consciousness researchers are engaging in a research project with an eye towards success and since, as I argue below, success carries with it certain ethical risks that would not arise if the research were not pursued, we should, perhaps artificially, assume that the risk of creating artificial consciousnesses is relatively high until we have reason to think otherwise.[20]

3.2 THE PROBABILITY OF MISTREATMENT

Whereas it is extremely difficult to predict how probable it is that a given research program will result in an artificial consciousness that is a moral patient, it is not so difficult to see why if a program were successful there would be a substantial chance that the created consciousness would be mistreated.

In arguing that research involving the use of cognitive enhancement technologies on non-human research subjects is morally problematic, I've raised the worry that for a variety of reasons, we might be more likely to mistreat research subjects than in traditional research contexts.[21] This is because cognitive enhancement research might alter research subjects in ways that aren't detectable and that can't be communicated by the research subjects themselves. Also, without further education about the concerns associated with cognitive enhancement, researchers might fail to take required precautions.

Similar worries arise with respect to artificial consciousnesses. Artificial consciousness research, unlike research involving non-human research subjects, is not subject to oversight designed to protect research subjects. Without oversight and researcher education, researchers are less likely to take the welfare of research subjects into account.

Furthermore, depending on which methods of creating artificial consciousness are successful, researchers may be in a poor epistemic situation with respect to determining whether they've created a moral patient at all. To see why, consider the following highly stylized case:

> **Selection**: Artificial consciousness researchers, informed by evolutionary biologists, have devised a series of problems that they think will encourage the evolution of consciousness. Programs are written that mutate with imperfect replication and reproduce proportionally to the efficacy with which they solve the various problems.

Let's say that the research program described in Selection is very likely to lead to programs that are conscious in ways that make those programs moral patients. It doesn't thereby follow that researchers will immediately know when such beings evolve or how to promote or avoid frustrating the interests of the created beings. Just because chimps and dogs are both moral patients, it doesn't thereby follow that treating them appropriately means treating them similarly. The same goes for artificial consciousnesses. Even if we become fairly confident that we've created an artificial consciousness, we can't be sure we know what is thereby required of us.

Of course, if it is impossible for researchers to determine what the interests of such research subjects are, they may be excused for any and all treatment that isn't commensurate with the moral patiency of these beings. But researchers are required to try to make determinations about the interests of these beings and to try to treat them appropriately in the context of doing research. What would be problematic in the above case would be for the researchers to experiment on such intelligences without making attempts to determine what's good for them.

3.3 THE CASE AGAINST CREATING ARTIFICIAL CONSCIOUSNESSES REVISITED

What does the case against creating artificial consciousnesses amount to? The above concerns provide a *pro tanto* case, or an overrideable presumption against artificial consciousness research which aims to create artificial consciousnesses that have capacities, such as sentience, self-awareness, or desires of one form or other, like mammals, birds, or humans. It is a presumption only, not an all things considered objection to such research. Furthermore, what is required to override the presumption depends in part on which normative theory is true, which will partially determine what is owed to moral patients.

What kinds of considerations will override the presumption against creating artificial consciousnesses? First, as research becomes more advanced, researchers might be able to determine that a particular research program is valuable but that the risk of creating a moral patient is extremely low. In such case, the ethical risk to the research subject is low and the presumption is overridden.

Second, researchers might be engaged in research that is likely to result in the creation of moral patients, but are taking or intend to take sufficient care to determine what constitutes appropriate treatment of the created research subjects. That is, they will not merely conduct their research, but also conduct research to determine what promotes or frustrates the interests of created beings. In doing so, the researchers lower the probability of mistreatment and the presumption is overridden.

Finally, it might be that artificial consciousness research is so valuable, and the cost of not doing it so costly, that it is worth doing no matter how poorly life goes for the entities created and where doing the research efficiently rules out taking the time to discern what's good or bad for research subjects. There may be an ethical imperative to engage in artificial consciousness research. And if so, and if doing so efficiently, requires that we ignore the interests of the created entities, then the presumption may be overridden. Just as with research involving non-human animals, the relevant question is whether the value of doing the research justifies the ethical costs accrued in harming research subjects. Sometimes, the answer is "yes."

However, whether the presumption is overridden for this reason is going to be extremely sensitive to the nature of our moral obligations, and, thereby, to which normative theory is ultimately true.[22] For a Utilitarian, it will be permissible to ignore the interests of artificial consciousnesses if doing so maximizes utility. The conditions under which a Kantian will agree to ignore or allow the interests of such patients to be overridden will be much more stringent.

4. CONCLUSION

The above is not meant to amount to a complete defense of the presumption against creating artificial consciousnesses, and, in fact, it leaves open how strong the ethical presumption is. Instead, I hope that it raises the ethical concerns associated with such research enough to start a conversation about the ethics of engaging in it. Ultimately, each artificial consciousness research program will have to be evaluated individually to assess the ethical risks, just as each research program involving non-human animals is evaluated. However, it is important that we recognize that there are ethical risks concerning the research subjects and not only those risks that accrue to us or that we face once we've realized artificial consciousness like our own.

NOTES

1. See, for example, Chalmers, "The Singularity."
2. See for example, O'Neill, "Varieties of Intrinsic Value"; Cahen, "Moral Considerability"; Sandler and Simons, "Artefactual Organisms." For dissent on the usefulness of moral status talk see, Sachs, "Moral Status."
3. Buchanan, *Beyond Humanity?*, chap. 7, for example, discusses the differences between moral status on a contractualist framework and moral status on a utilitarian framework. See also Sober, "Philosophical Problems for Environmentalism."
4. Another version of the Dependency Thesis might claim that the degree to which a being has moral status depends on normative theory. Buchanan, in *Beyond Humanity?*, seems to suggest this as well. However, I think this version of dependency is also false. There are ways to cash out differences in treatment owed to different kinds of beings without understanding them as having different degrees of moral status. In other words, "degrees of moral status" can be gotten rid of without losing the ability to make the normative distinctions that talk is intended to capture. This translatability is not central to what I'll say here and so I leave it unargued for.
5. Consider, for example, the difference between Singer's view about the moral status of humans and Frey's view of same. Both are committed utilitarians and yet Singer (*Animal Liberation*) thinks that all sentient beings are equal, that is have equal moral status (though Singer acknowledges that typically, a human's life should often be preferred over an animal's in a conflict because humans can suffer and enjoy in more ways than most non-human animals) while Frey ("Vivisection"; "Moral Standing") thinks that human adults of sound mind are distinct from non-human animals, that their lives are of more value because of their capacity for certain kinds of experiments. It is worth noting that both come to similar

conclusions about the ethics of animal experimentation and the differences between their views are subtle, but the fact that Frey thinks humans have additional value in virtue of having a capacity or capacities that non-human animals do not is sufficient to demonstrate the kind of inter-normative differences in conceptions of moral status that are relevant here.

6. See, for example, Regan, *Case for Animal Rights*, who argues, using arguments very similar to those employed by Singer, to extend a Kantian conception of rights to non-human animals that are minimally conscious. See also Rollin, *Animal Rights & Human Morality*, for a discussion that includes contractualist discussions of animal moral status. For an excellent discussion of how a more traditional Kantian might approach the issue of animal rights, see Korsgaard, "Fellow Creatures."

7. Of course, which properties are taken to be morally significant are often influenced by which normative theory one takes to be true. A Kantian is more likely to think that "being an end in oneself" is a morally significant property than a utilitarian. But, that is a sociological fact. The Kantian still owes the utilitarian an argument as to why that property is morally significant. If the argument is sounds, the utilitarian might agree that it is only the benefits and harms that accrue to ends in themselves that influence the value of states of affairs, just as many utilitarians are keen to think that it is only the benefitting and harming of humans that make a difference to the value of states of affairs.

8. We could of course understand a normative theory to include facts about whom or what has moral status. I'm using normative theory, as is typical, to pick out a theory of right action (and, if you like, an account of the source of normativity).

9. For a more detailed explanation, see Basl, "Machines as Moral Patients."

10. For an overview of these families, see Griffin, *Well-Being*; Streiffer and Basl, "Applications of Biotechnology."

11. I'm not here committing to the view that my dog's welfare matters for its own sake simply because she has a welfare. It might be that her welfare matters because she is an end in herself, or because reasonable would agree that an animal's welfare is morally significant. Again, I'm not committing to any particular normative theory or any particular source of normativity. Whichever theory is true, I explain below, my dog's welfare is relevant to moral deliberations for her own sake.

12. See, for example, Feinberg, "Rights of Animals," on plants.

13. This isn't to say how their welfare affects our own or others isn't also relevant to deliberations. In thinking about what to do, we must think about these conflicts of interests. That is consistent with thinking that a being's interests should be taken into account for the sake of the being under consideration.

14. Lewis, *Plurality of Worlds*.

15. For a discussion of the distinction between obligation and excuse, see McMahan, *Killing in War*.

16. Basl, "Machines as Moral Patients."

17. See Shiffrin, "Wrongful Life," for a dissenting argument that bringing a child into existence is a *pro tanto* wrong.

18. Streiffer, "Edge of Humanity."

19. Basl, "Sensitivity Enhancement."

20. Some projects that might be classified as "artificial consciousness projects" but are thought to involve only preliminary research or are being done as development steps should be excluded from the scope of this assumption.

21. Basl, "Sensitivity Enhancement."

22. The same can be said of traditional research. While some Kantians, for example, endorse a complete prohibition on any animal research, most recognize that the rights of all beings are not, in fact, inviolable come what may. If circumstances are such that great harms can't be avoided, some individuals may be sacrificed for others. Regan, *Case for Animal Rights*; Buchanan, *Beyond Humanity?*

BIBLIOGRAPHY

Basl, John. "Machines as Moral Patients We Shouldn't Care About (Yet)." *Philosophy and Technology*, forthcoming.

———. "Sensitivity Enhancement: The Ethics of Engineering Non-Human Research Subjects." In *Designer Biology: The Ethics of Intensively Engineering Biological and Ecological Systems*, 219–232. Lexington Books, 2013.

Buchanan, Allen E. *Beyond Humanity?: The Ethics of Biomedical Enhancement*. Oxford University Press, USA, 2011.

Cahen, Harley. "Against the Moral Considerability of Ecosystems." In *Environmental Ethics: An Anthology*, edited by Andrew Light and H. Rolston III. Wiley-Blackwell, 2002.

Chalmers, David. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17, no. 9-10 (2010): 7–65.

Feinberg, Joel. "The Rights of Animals and Future Generations." *Columbia Law Review* 63 (1963): 673.

Frey, R. G. "Vivisection, Morals and Medicine." *Journal of Medical Ethics* 9, no. 2 (1983): 94.

———. "Moral Standing, the Value of Lives, and Speciesism." *Between the Species: a Journal of Ethics* 4, no. 3 (1988): 191.

Griffin, James. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford University Press, 1988.

Korsgaard, Christine M. "Fellow Creatures: Kantian Ethics and Our Duties to Animals." *The Tanner Lectures on Human Values* 25 (2004): 26.

Lewis, David K. *On the Plurality of Worlds*. Malden, Mass.: Blackwell Publishers, 2001.

McMahan, Jeff. *Killing in War*, 1st ed. Oxford: Oxford University Press, 2009.

O'Neill, John. "The Varieties of Intrinsic Value." In *Environmental Ethics: An Anthology*, edited by Holmes III Rolston and Andrew Light. Wiley-Blackwell, 2003.

Regan, Tom. *The Case for Animal Rights*. Berkeley: University of California Press, 1983.

Rollin, Bernard E. *Animal Rights & Human Morality*. Amherst, NY: Prometheus Books, 2006.

Sachs, Benjamin. "The Status of Moral Status." *Pacific Philosophical Quarterly* 92, no. 1 (2011): 87–104.

Sandler, Ronald, and Luke Simons. "The Value of Artefactual Organisms." *Environmental Values* 21, no. 1 (2012): 43–61.

Shiffrin, S. V. "Wrongful Life, Procreative Responsibility, and the Significance of Harm." *Legal Theory* 5, no. 2 (1999): 117–148.

Singer, Peter. *Animal Liberation*, vol. 1. New York: Ecco, 2002.

Sober, Elliott. "Philosophical Problems for Environmentalism." In *The Preservation of Species*, edited by Bryan Norton. Princeton University Press, 1986.

Streiffer, Robert. "At the Edge of Humanity." *Kennedy Institute of Ethics Journal* 15, no. 4 (2005): 347–70.

Streiffer, Robert, and John Basl. "Applications of Biotechnology to Animals in Agriculture." In *The Oxford Handbook of Animal Ethics*, edited by T. Beauchamp and R. Frey. Oxford: Oxford University Press, 2011.

Turing Test, Chinese Room Argument, Symbol Grounding Problem: Meanings in Artificial Agents

Christophe Menant

INDEPENDENT SCHOLAR

HTTP://CRMENANT.FREE.FR/HOME-PAGE/INDEX.HTM

ABSTRACT

The Turing Test (TT), the Chinese Room Argument (CRA), and the Symbol Grounding Problem (SGP) are about the question "can machines think?" We propose to look at these approaches to Artificial Intelligence (AI) by showing that they all address the possibility for Artificial Agents (AAs) to generate meaningful information (meanings) as we humans do. The initial question about thinking machines is then reformulated into "can AAs generate meanings like humans do?"

We correspondingly present the TT, the CRA, and the SGP as being about generation of human-like meanings. We model and address such possibility by using the Meaning Generator System (MGS) where a system submitted to an internal constraint generates a meaning in order to satisfy the constraint. The system approach of the MGS allows comparing meaning generations in animals, humans, and AAs. The comparison shows that in order to have AAs capable of generating human-like meanings, we need the AAs to carry human constraints. And transferring human constraints to AAs raises concerns coming from the unknown natures of life and human mind which are at the root of human constraints. Implications for the TT, the CRA and the SGP are highlighted. It is shown that designing AAs capable of thinking like humans needs an understanding about the natures of life and human mind that we do not have today. Following an evolutionary approach, we propose as a first entry point an investigation about the possibility for extending a "stay alive" constraint into AAs. Ethical concerns are raised from the relations between human constraints and human values. Continuations are proposed. (This paper is an extended version of the proceedings of an AISB/IACAP 2012 presentation (http://www.mrtc.mdh.se/~gdc/work/AISB-IACAP-2012/NaturalComputingProceedings-2012-06-22.pdf).

1. TURING TEST, CHINESE ROOM ARGUMENT, AND MEANING GENERATION

The question "can machines think?" was addressed in 1950 by Alan Turing and formalized by a test, the Turing Test (TT), where a computer is to answer questions asked by humans. If the answers coming from the computer are not distinguishable from the ones made by humans, the computer passes the TT.[1] So the TT addresses the capability for a computer to understand questions formulated in human language and answer these questions as well as humans would do. Regarding human language, we consider that understanding a question is to access the meaning of the question. And answering a question obviously goes with generating the meaning of the answer. So we consider that the TT is about meaning generation.

The validity of the TT was challenged in 1980 by John Searle with a thought experience, the Chinese Room Argument (CRA), aimed at showing that a computer can pass the TT without understanding symbols.[2] A person not speaking Chinese and exchanging Chinese symbols with people speaking Chinese can make them believe she speaks Chinese if she chooses the symbols by following precise rules written by Chinese speaking persons. The person not speaking Chinese passes the TT. A computer following the same precise rules would also pass the TT. In both cases the meaning of the Chinese symbols is not understood. The CRA argues that the TT is not valid for testing machine thinking capability as it can be passed without associating any meaning to the exchanged information. Here also, the understanding of the symbols goes with generating the meanings related to the symbols. So we can consider that the TT and the CRA are about the possibility for AAs to generate human-like meanings. This turns the question about whether machines are capable of thinking into a question on meaning generation. Can AAs generate human-like meanings?

In order to compare the meanings generated by humans and by AAs, we use the Meaning Generator System (MGS). The MGS models a system submitted to an internal constraint that generates a meaning when it receives information that has a connection with the constraint. The generated meaning is precisely the connection existing between the received information and the constraint, and it is used to determine an action that will be implemented in order to satisfy the constraint.[3]

The MGS is simple. It can model meaning generation in elementary life. A paramecium moving away from acid water can be modeled as a system submitted to a "stay alive" constraint that senses acid water and generates a meaning "presence of acid not compatible with the 'stay alive' constraint." That meaning is used to trigger an action from the paramecium: get away from acid water. It is clear that the paramecium does not possess an information processing system that would allow her to have access to an inner language. But a paramecium has usage of sensors that can participate in a measurement of the acidity of the environment. The information made available with the help of these sensors will be part of the process that will generate the move of the paramecium in the direction of less acid water. So we can say that the paramecium has overall created a meaning related to the hostility of her environment in connection with the satisfaction of her vital constraint. Figure 1 represents the MGS with this example.

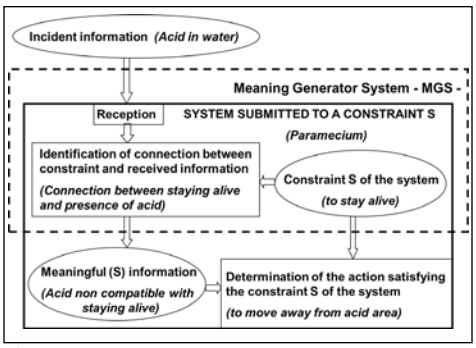

Figure 1. The Meaning Generator System.

The MGS is a simple tool modeling a system submitted to an internal constraint.[4] It can be used as a building block for higher level systems (agents) like animals, humans, or AAs, assuming we identify clearly enough the constraints corresponding to each case.[5]

The function of the meaningful information is to participate in the determination of an action that will be implemented in order to satisfy the constraint of the system. This makes clear that a meaning does not exist by itself. A meaning is meaningful information about an entity of the environment which is generated by and for a system submitted to an internal constraint that characterizes the system.

The MGS approach is close to a simplified version of the triadic Peircean theory of sign (Sign, Object, Interpretant). Peirce's theory is a general theory of sign and the MGS approach is centered on meaning. The MGS can be compared to a simplified version of the Peircean Interpreter producing the Interpretant. The generated meaning combines an objective

entity of the environment (the incident information) and a specific construction of the system (the connection with the constraint). The MGS displays a simple complementarity between objectivism and constructivism.

The MGS is also usable to position meaning generation in an evolutionary approach. The starting point is basic life with a "stay alive" constraint (for individuals and for species) and a "group life" constraint. The sight of a cat generates a meaning within a mouse, as well as a passing by fly within a hungry frog. But the "stay alive" constraint refers to life, the nature of which is unknown as of today. What can be accessed and analyzed are the actions that will be implemented to satisfy the "stay alive" constraint, not the constraint. For humans, the constraints are more difficult to identify. They are linked to human consciousness and free will, which are both mysterious concepts for today's science and philosophy. However, some aspects of human constraints are however easy to guess, like "look for happiness" or "limit anxiety."[6] References to the Maslow pyramid can also be used as an approach to human constraints.[7] But what can be understood about these constraints refers mostly to the actions implemented to satisfy them. The nature of the constraints is unknown as related to the still mysterious human mind.

In all cases the action implemented to satisfy the constraint will modify the environment, and so the generated meaning. As said, meanings do not exist by themselves. They are agent related and come from meaning generation processes that link the agents to their environments in a dynamic mode. Different systems can generate different meanings when receiving the same information. And incident information can be meaningful or meaningless.[8]

Most of the time agents contain several MGSs related to different sensorimotor systems and different constraints to be satisfied. An item of the environment generates different interdependent meanings that build up networks of meanings representing the item to the agent. These meaningful representations embed the agent in its environment through constraints satisfaction processes.

To see if AAs can generate meanings like humans do we have to look at how human meaning generation processes could be transferred to AAs. Figure 1 shows that the constraint is the key element to be considered in the MGS. The other elements deal with data processing that is transferrable. But when looking at transferring human constraints to AAs, we face the problem of the unknown natures of life and human mind from which these constraints result. Take for instance the basic "stay alive" constraint that we share with animals. We know the actions that are to be implemented in order to satisfy that constraint, like keep healthy and avoid dangers. But we do not really know what life is. We understand that life came out of matter during evolution, but we do not know how life could be today built up from inanimate matter. The nature of life is a mystery. Consequently, we cannot transfer a "stay alive" constraint to AAs because we cannot transfer something we do not understand. The same applies for human specific constraints which are closely linked to human mind. We do not know exactly what is "look for happiness." We only know (more or less) the physical or mental actions that should be implemented in order to satisfy this complex constraint. So we have to face the fact that the transfer of human constraints to AAs is not today possible as we cannot transfer things we do not know.

The proposed approach shows that we cannot today build AAs able to generate human-like meanings. In the TT, the computer is not in a position to generate meanings like humans do. The computer cannot understand the questions nor the answers as humans do. It cannot pass the TT. Consequently, the CRA is right. Today AAs cannot think like humans think. Strong AI is not possible today. A better understanding about the natures of life and human mind is necessary for a progress toward the design of AAs capable of thinking like humans think. Research activities are in process in these areas.[9] Some possible short cuts may be investigated, at least for the transfer of animal constraints (see hereunder).

2. SYMBOL GROUNDING PROBLEM AND MEANING GENERATION

The possibility for computers to attribute meanings to words or symbols has been formalized by Stevan Harnad in 1990 through the Symbol Grounding Problem (SGP).[10] The SGP is generally understood as being about how an AA computing with meaningless symbols can generate meanings that are intrinsic to the AA. "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?"

The SGP being about the possibility for AAs to attribute intrinsic meanings to words or symbols, we can use the MGS as a tool for an analysis of the intrinsic aspect of the generated meaning. The MGS defines a meaning for a system submitted to an internal constraint as being the connection existing between the constraint and the information received from the environment. The intrinsic aspect of the generated meaning results from the intrinsicness of the constraint. In order to generate an intrinsic meaning, an agent has to be submitted to an intrinsic constraint. Putting aside metaphysical perspectives, we can say that the performance of meaningful information generation appeared on earth with the first living entities. Life is submitted to an intrinsic and local "stay alive" constraint that exists only where life is and is not present in the material world surrounding the living entity. As today AAs are made with material elements, they cannot generate intrinsic meanings because they do not contain intrinsic constraints. So the semantic interpretation of meaningless symbols cannot be intrinsic to AAs. The SGP cannot have a solution in the world of today AAs.[11]

The same conclusion can be reached by recalling the impossibility to transfer human constraints to AAs. The constraints that are present in the AAs are derived constraints implemented by the designers (like win chess or avoid obstacles). These constraints come from the designer of the AA. They are not intrinsic to the agent like are "stay alive" or "look for happiness" constraints. AAs can only generate derived meanings coming from their derived constraints. Today AAs cannot carry intrinsic constraints and consequently

cannot generate intrinsic meanings. Again, the SGP cannot have a solution in the world of today's AAs.

The conclusions reached in the previous paragraph apply. AAs cannot today generate meanings nor think like we humans do. We need better understandings about the natures of life and human mind in order to address the possibility for human-like meaning generation and thinking in AAs.

Another area of investigation for intrinsic constraints in AAs is to look for AAs capable of creating their own constraints. Whatever the possible paths in this area, it should be highlighted that such approach would not be enough to allow the design of AAs able to think like humans do. The constraints that the AAs might be able to generate by themselves may be different from human ones or managed differently by the AAs. These future AAs may think, but not like humans think. This brings up ethical concerns for AI where AAs would not be managing constraints and meanings the same way humans do.

3. ARTIFICIAL INTELLIGENCE, ARTIFICIAL LIFE, AND MEANING GENERATION

The above usage of the MGS with the TT, the CRA, and the SGP has shown that machines cannot today think like humans do because human constraints are not transferrable to AAs. The basic "stay alive" constraint is also part of human constraints, and not being able to transfer it to AAs implies that we cannot design AAs managing meanings like living entities do. Strong artificial life (AL) is not possible. So not only can't we design AAs able to think like humans think, we can't even design AAs able to live like animals live. At this level of analysis, the blocking points in AI and in AL come more from our lack of understanding about the natures of life and human mind than from a lack of computer performances. Progresses in AL and in AI need more investigations about the nature of life and the nature of human mind.

In terms of increasing complexity, these subjects can be positioned following an evolutionary approach. As life came up on earth before human mind, it should be easier and logical to address first the problem about the "stay alive" constraint not transferrable to AAs. Even if we do not know the nature of life, we are able to manipulate it. And we could, instead of trying to transfer the performances of life to AAs, look at how it could be possible to extend life to AAs, without needing an understanding about the nature of life. In a way to be defined, we would bring the AA at the level of a living entity. We would design an agent being at the same time alive and artificial. An agent being alive (submitted to a "stay alive" constraint), and being artificial (on which we keep some control). Research activities are in process on close domains like integrating the computational capabilities of neurons in robots control circuits or designing insect-machine hybrids with motor control of insects.[12] These research activities are promising for the development of biological computing and life-AA merging, but the possibility for extending a "stay alive" constraint to the AA is to be investigated. Such possible progress about having AAs submitted to resident animal constraints does not solve the problem of AAs submitted to human constraints, but we can take this as a first step in an evolutionary approach to AAs containing human constraints.

4. MEANING GENERATION, CONSTRAINTS, VALUES, AND ETHICAL CONCERNS

The MGS approach has shown that our current lack of understanding about the natures of life and human mind make impossible today the design of AAs able to think like humans do. The reason being that we do not know how to transfer human constraints (like "look for happiness") to AAs. But human constraints do not a priori include human values (some humans find happiness by the suffering of others). So looking at transferring human constraint to AAs brings up ethical concerns. Artificial agents submitted to human constraints may not carry human values. Research on the nature of human mind and artificial intelligence should consider how human values could be linked to human constraints. It is a challenging subject because human values are not universal and human constraints remain ill defined. But the nature of human mind is still to be discovered and we can hope that its understanding will shed some light on the diversity of human values. Also, as addressed above, another case is the one about AAs becoming capable of generating by themselves their own constraints. Such an approach should keep human values in the background of these constraints so the AAs are not brought to generate meanings and actions too distant from human values.

5. CONCLUSIONS

We have proposed that the TT, the CRA, and the SGP can be understood as being about the possibility of AAs generating human-like meanings. Using that analogy, it has been argued that AAs cannot think like humans think because they cannot generate human-like meanings. This has been shown by using a model of meaning generation for internal constraint satisfaction (the MGS). The model shows that our lack of understanding about the natures of life and human mind makes impossible the transfer of human constraints to AAs. Consequently, today AAs cannot think like we humans think. They cannot pass the TT. The CRA is correct and the SGP cannot have a solution. Strong AI is not possible today. Only weak AI is possible. Imitation performances can be almost perfect and make us believe that AAs generates human-like meanings, but there is no such meaning generation as AAs do not carry human constraints. Artificial agents do not think like we do. Another consequence is that it is not possible today to design living machines. Artificial agents cannot generate meanings like animals do because we do not know the nature of life and cannot transfer animal constraints to AAs. Strong AL is not possible today. At this level of analysis the blocking points for strong AI and strong AL come more from our lack of understanding about life and human mind than from computers performances. We need progress in these understandings to design AAs capable of behaving like animals and thinking like humans. As life is less complex and easier to understand than consciousness, the transfer of a "stay alive" constraint to AAs should be addressed first. An option could be to extend life with its "stay alive" constraint within AAs. The AA would then be submitted to the constraints brought in with the living entity.

Ethical concerns have been raised through the possible relations between human constraints and human values. If AAs can someday be submitted to human constraints, they may not carry human values.

6. CONTINUATIONS

The MGS approach applied to the TT, the CRA, and the SGP has shown that the constraints to be satisfied are at the core of the meaning generation process and that it is not possible today to transfer animal or human constraints to AAs because of our lack of understanding about life and human mind. As a consequence it is not possible today to design AAs that can live like animals or think like humans. This status leads us to consider further developments linking constraints, life, and human mind in an evolutionary background.

An evolutionary approach to the nature of constraints should open the way to an understanding of a continuity of constraints from animal to humans. It would support an evolutionary theory of meaning and may provide new perspectives for an understanding about the nature of life and the nature of human mind. It may also support the possibility of addressing human constraints without using animal ones (i.e., addressing strong AI without usage of strong AL).

Identifying the origin of biological constraints relatively to physico-chemical laws may allow us to start an evolutionary theory of meaning in the material world. Work is in process on these subjects.[13]

The MGS approach also offers the possibility of defining meaningful representations that embed agents in their environments. Such representations can be used as tools in an evolutionary approach to self-consciousness where the human constraints play a key role. Work is in process in this area.[14]

An evolutionary approach to human constraints would address the "stay alive" constraint that we share with animals. But the nature of life is a mystery today. As introduced above, we feel it could be interesting to investigate the possibility of having a living entity extend its "stay alive" constraint within AAs. We could then have AAs submitted to the "stay alive" constraint without needing an understanding about the nature of life.

Regarding ethical concerns, an evolutionary approach to human consciousness could introduce a common evolutionary background for constraints and values. Such concern applies also to the possibility of AAs creating their own constraints that may be different from human ones and consequently not linked to human values.

NOTES

1. Turing, "Computing Machinery."
2. Searle, "Minds, Brains and Programs."
3. Menant, "Information and Meaning."
4. In the MGS approach the constraint is proper to the system that generates the meaning (see Figure 1). The constraint is related to the nature of the system.
5. The MGS approach is based on meaning generation for constraint satisfaction. It is different from "action oriented meaning." With the MGS, the constraint to be satisfied is the cause of the generated meaning which determines the action that will be implemented to satisfy the constraint. The meaning is then "constraint satisfaction oriented." The action comes after Menant, "Computation on Information, Meaning and Representations."
6. "Anxiety limitation" has been proposed as a constraint feeding an evolutionary engine that could have lead pre-human primates to the performance of self-consciousness. Menant, "Information and Meaning in Life, Humans and Robots"; Menant, "Evolution and Mirror Neurons"; Menant, "Evolution and Mirror Neurons."
7. Menant, "Computation on Information, Meaning and Representations."
8. Such usage of meaningful information is different from the Standard Definition of Semantic Information (SDI) linked to linguistics where information is meaningful data (Floridi, "From Data to Semantic Information"). Our system approach addresses all types of meaning generation by a system submitted to an internal constraint. It covers the cases of non linguistic meanings (animals and AAs).
9. Philpapers, "Nature of Consciousness" search results; Philpapers, "Nature of Life" search results.
10. Harnad, "Symbol Grounding Problem."
11. Several proposals have been made as solutions to the SGP. Most have been recognized as not providing valid solutions. Taddeo and Floridi, "Solving the Symbol Grounding Problem."
12. Warwick et al., "Controlling a Mobile Robot"; Bozkurt et al., "Insect–Machine Interface Based Neurocybernetics."
13. Riofrio, "Informational Dynamic Systems."
14. Menant, "Evolutionary Advantages."

BIBLIOGRAPHY

Bozkurt, Alper, R. F. Gilmour, Ayesa Sinha, David Stern, and Amit Lal. "Insect–Machine Interface Based Neurocybernetics." *Biomedical Engineering, IEEE Transactions on*. 56, no. 2 (2009): 1727–33.

Floridi, Luciano. "From Data to Semantic Information." *Entropy* 5 (2003): 125–45. http://mdpi.muni.cz/entropy/papers/e5020125.pdf.

Harnad, Stevan. "The Symbol Grounding Problem." *Physica D* 42 (1990): 335–46.

Menant, Christophe. "Information and Meaning." *Entropy* 5 (2003): 193–204. http://mdpi.muni.cz/entropy/papers/e5020193.pdf.

———. "Information and Meaning in Life, Humans and Robots." *Proc. of the 3rd Conference on the Foundation of Information Sciences*, Paris, 2005a.

———. "Evolution and Mirror Neurons. An Introduction to the Nature of Self-Consciousness." *TSC 2005*. Copenhagen, Denmark 2005b. http://cogprints.org/4533/.

———. "Evolutionary Advantages of Inter-subjectivity and Self-Consciousness through Improvements of Action Programs." *TSC 2010*. Tucson, AZ, 2010. http://cogprints.org/6831/.

———. "Computation on Information, Meaning and Representations. An Evolutionary Approach." In *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*, edited by G. Dodig-Crnkovic and M. Burgin, 255–86. World Scientific, 2011.

Philpapers. "Nature of Consciousness." Search results. 2013. http://philpapers.org/s/nature%20of%20consciousness.

Philpapers. "Nature of Life." Search results. 2013. http://philpapers.org/s/nature%20of%20life.

Riofrio, Walter. "Informational Dynamic Systems: Autonomy, Information, Function." In *Worldviews, Science, and Us: Philosophy and Complexity*, edited by Carlos Gershenson, Diederik Aerts, and Bruce Edmonds, 232–49. Singapore: World Scientific, 2007.

Searle, John R. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3 (1980): 417–24.

Taddeo, Mariarosario, and Luciano Floridi. "Solving the Symbol Grounding Problem: Critical Review of Fifteen Years of Research." *Journal of Experimental & Theoretical Artificial Intelligence* 17, no. (2005).

Turing, Alan M. "Computing Machinery and Intelligence." *Mind* 59 (1950): 433–60.

Warwick, Kevin, Dimitris Xydas, Slawomir J. Nasuto, Victor M. Becerra, Mark W. Hammond, Julia H. Downes, Simon Marshall, and Benjamin J. Whalley. "Controlling a Mobile Robot with a Biological Brain." *Defence Science Journal* 60, no. 1 (2010): 5–14.

Assistive Environment: The Why and What

Linda Sebek
MÄLARDALENS HÖGSKOLA SCHOOL OF INNOVATION, DESIGN, AND ENGINEERING

INTRODUCTION

Assistive environments are coming strong, taking advantage of progress in robotics and embedded technology. It is not hard to picture a future with ubiquitous computers taking care of everything we find tedious, but the same technologies can be used to give everyone the same possibilities to live their life their own way.[1]

Today a lot of people with impairments depend on other people for help in their daily lives. This support enables them to live normal lives, but this dependence can also hinder them to do as they like. Human helpers have a tendency to interfere, to have opinions about how the impaired person should do and live, just because they are humans. They want to help, but also to protect and take care.[2]

Technology is already used to provide necessary support; for example, a computerized communication device such as a PDA[3] will speak out the words you want to say, but are unable to. Before you had to rely on a human to interpret your feelings and wishes and translate them into words. The communication device will not interpret, it will only translate. It will become your voice.

There are several experimental homes with smart assistive environments that will support the occupants in the house. Most of these target the elderly in an attempt to support them so that they can stay in their own homes longer. Many services provided relate to health and supervision of health.[4] Though this might be important even to people with impairments, there might be other, more urgent, areas to target.

This article will argue for the need for more research to find out what functions an assistive environment must be able to perform, and to identify the needs that must be met.

IMPAIRMENT AND DISABILITY

Impairment is caused by a bodily dysfunction, such as movement impairment. Disability, on the other hand, can have several different meanings. Traditionally, disability is considered to be synonymous with impairment. Disability then is a disadvantage caused by the person's body. The social model of disability is in opposition to this definition and sees disability as something outside the person. Thomas has a very good way of describing it: "disability comes into being when aspects of contemporary social structure and practice operate to disadvantage and exclude people with impairments through restrictions of their activity."[5]

So disability is really an effect of disablism in our society—not an effect of personal impairments.[6] The social model is about empowering those who need the power, people with impairments. Society needs to strive to reduce the need for control of outsiders over the lives of people with impairments by reducing the power that professionals have over the support people with impairments are dependent on.[7]

Disability can also be viewed as an administrative category. In a welfare state less fortunate people get institutional support. Disability is a category that, in the eye of the welfare system, is a privileged group, due to the fact that people belonging to the category will get support in some way. The laws regulating this differ from country to country. As the welfare state must be sure that the right people get support, it becomes important to group people as "disabled" and "non-disabled." Whether you are viewed as disabled or non-disabled depends on if your impairment fits in the category. With an administrative definition on disability, those that get disability benefits are disabled.[8]

INTELLECTUAL IMPAIRMENTS

This article uses the term intellectual impairments over mental retardation, mentally disability, intellectual disability, or learning disability. Mental retardation is an old term that many people find offensive and as a result it should not be used. The other terms all include disability and, as earlier explained, disability can mean different things.

Switsky and Greenspan define intellectual impairments as a dysfunction of intellectual, social and in adaptive skills.[9] Adaptive skills concern how well a person functions in everyday life when it comes to communication, hygiene, health, keeping safe, and more. The adaptive skill affects many different areas in a person's life and is a determining factor for what support the person needs.

An environment better adapted to a person will actually influence the impact low adaptive skills will have on a person's life.

ASSISTIVE ENVIRONMENTS

Helal, Lee, and Mann define an assistive environment as "a smart environment specifically designed to serve individuals with special needs, including older people and people with disabilities."[10]

I suggest that the assistive environment is an intelligent environment designed to serve every human, no matter the individual human's capability. In a home, this must be customized to the occupants of the home. In the workplace, the environment must be adapted to meet the needs of the workforce and customers. Assistive environments should not be limited to certain groups as that raises a need to define those groups and classify people in order to decide who belongs to the group and who doesn't. Another way to address this is to develop the technology with the capability to be upgraded easily. A basic platform for intelligent environment could even be made standard in every house, just as an electric system is standard today. This way a personal adaptation can be made to anyone whether the need of assistance is permanent or temporary. Baldoni et al. are looking into ways to develop a middleware platform that looks promising for building smart homes for everyone.[11]

In an intelligent environment, computers are connected with everyday tasks and it becomes important that the system is partly self-regulating. It needs to be able to adjust to new conditions and communicate with residents in a way that does not disturb everyday life. It is not people who must adapt to the machines, but the machines must be designed to fit into the life of humans.[12]

Specialized computers used in machines to control things like home equipment are embedded systems. If we want an environment that interacts with people then the embedded systems will play an important part. We would want the environment to give us information whenever asked for it and to respond to other requests from us. There are still plenty of technical problems to deal with, but this article leaves those to the engineers to solve.[13]

Embedded computing available everywhere to assist in everyday life is called ubiquitous computing. The ubiquitous computing plays an important part in creating an assistive environment that will support all humans.[14]

The environment's ability to read, understand, and appropriately respond to the information/data from the environment's sensor is an important aspect of the assistive environment. This ability of apt response determines the benefits for the user as the environment can provide more appropriate service.[15]

An interesting version is the assisted cognition system that is designed to learn when to offer help to occupants in intelligent environments. The system comes to understand everyday behavior patterns and when assistance is needed. If the occupant is starting to go in the wrong direction when walking from the store and home, a handheld device might display an arrow pointing in the right direction.[16] Here, of course, it is very important that the engineering solution offering guidance does not impose the command over the human. Again, here is a problem for an engineer in collaboration with psychologists, medical personnel, and, most importantly, the targeted users. In a way it is similar to the GPS system, which also offers assistance and may in principle make errors dangerous for the user.

FUNCTIONS NEEDED

Not much research has been done about what kind of support people with intellectual impairments need for independent living. Ringsby-Jansson has looked at different supporting facilities for living in several communities in Sweden. She mentions only a little about the actual kind of support the occupants get in these facilities. The support is delivered by human helpers and includes help with cleaning the house, laundry, and cooking. The residents also need help with structuring their days and planning their economy. She mentions one occupant who likes to visit cafés, but he can only do this once a week when he has a helper that can join him.[17] Gough & Andersson also show that cleaning the living spaces is a burden for the staff, taking up a lot of their time. Buying and preparing food is one of the everyday tasks that many people need assistance with.[18]

CASE STUDY: SIMON

To gain further insight, I interviewed a mother of a teenager, Simon, with moderate intellectual impairment. What kind of support would Simon need to be able to move to his own apartment?

Simon has good learning capabilities when it comes to routine work, as long as he gets sufficient time to practice and learn how it is done. This means he needs help in all new tasks he meets, no matter if it's laundry or finding his way in new surroundings. Simon does not understand the impact junk food will have on his well-being. He wants to have desserts for dinner. He needs support to be able to choose and plan food that he both likes and is good for him. Simon can't read, but he knows the recipes for his favorite cakes by heart. If he notices an ingredient is missing, he will ask that it be bought, but he cannot write a list of things to buy. Not being able to read also presents a problem at the actual store; he will have a hard time finding what he needs if the store is big and has a lot to offer. Another reading-related area he needs help with is the mail. The mail might contain important messages such as hospital appointments, as well as junk mail. He needs help to identify important from unimportant, and help with reading and understanding if the mail is not written in a simple way.

Time and money are two other areas of concern. Simon has limited understanding of both, requiring assistance to plan his day so he won't miss out on activities he wants and needs to participate in. This also involves ongoing reminders to stay on task so that distractions won't make him late. If he is supposed to get his coat and go out to catch the bus, he does not understand that watching DVD-movies will make him miss the bus. He needs help managing his time to allow for both work and pleasure. Additionally, Simon has no concept of the value of money. He therefore needs help with managing his budget and in the actual handling of money. Simon also has a limited capability to determine whether something is clean or dirty. He needs help to know when it's time to wipe the floor or put a sweater in the laundry.

A FUTURE HOME FOR SIMON

Simon is a young man who would benefit greatly by an assistive living environment. I will now welcome you into what could be his smart home in the future.

The day starts with a wake-up call when it is time to get up, accompanied by information about the day's activities. If Simon does not get up, he will get the information again until sensors register he is up. At breakfast Simon gets a nutritious menu to choose from. He chooses buttermilk and granola. As he takes the last of the buttermilk he shows the package to the kitchen screen to put it on his grocery list. After breakfast he should be getting ready for work, but a magazine has caught his attention. Sensors now register that Simon is not in the hallway putting on his jacket in time and he gets a reminder. Simon goes to work and on his way home he stays at the store to get groceries. His PDA brings him the grocery list and guides him through the store. He also wants to go get a new pair of jeans. He is apprised of whether he has enough money in his account and where to go to find a store. At home, his robot vacuum cleaner has already done its job and the floors are tidy. He brings in the mail and it is read out loud to him. There is a reminder of an appointment with the dentist in the mail and this is automatically registered. In the evening a friend drops by and they decide they want to see a movie that will have a premiere later this week. Simon receives information on when the movie is showing and he is free to book tickets. Later he will be informed when it's time to go to bed so that he will not be too tired in the morning. As he takes his clothes off, he is informed that his shirt needs to go to the laundry basket.

Simon needs assistance that will overrule his immediate impulses, making sure he gets up, gets to work, and so

forth. Today this assistance is given by his mother, and in a regular assisted living facility there would be employees doing this. In order to empower Simon himself in a smart assistive environment he would need to be included in decision making. Every week he would sit down with a personal assistant and go through the week. She would listen to him and guide him and together they would decide what assistance Simon would receive in the coming week.

ETHICS AND ASSISTIVE ENVIRONMENTS

Assistive environments allow for rich, independent living, aided by machines that will obey rather than human aids that tend to have their own agenda. But the assistive environment also opens up risks for severe privacy violations. The ethics around privacy issues are therefore of utmost importance in the development of assistive environments. Babbit et al. (2006) suggest fair informational practices in their discussion about intelligent housing, which include:

- Openness and transparency – the occupant knows what is recorded, how it is used, and why.
- Individual participation and consent – the occupants have influence over what is recorded and have the power to change wrong information.
- Collection limitation and data quality – only information needed for a specific purpose is collected.
- User limitation – information is only to be used for the purpose it was collected for and is only to be handled by authorized individuals.
- Security – the level of security is in proportion to the sensitivity of the data.[19]

As many individuals with intellectual impairments often have contact with many different people through community services and health services, user limitation and security is extremely important. Only a few people who actually need access should have it. Less sensitive information might be more widely available, but everyone with even limited access must be trained when to access the information and, more importantly, when not to and how to handle sensitive information.

CONCLUSION

The case study is given as an illustration of a real-life situation of an impaired teenager. The support Simon needs can be traced back to low adaptive skills. An assistive environment that compensates for this shortcoming would give Simon more power over his own life. It is not possible to draw any broad conclusions from one single case, but it is clear that there are many needs beyond purely medical areas that a high-tech assistive environment can support. The Simon case shows us a few of these areas of daily life. More research into the types of support people with intellectual impairments need for independent living will also tell us what functions an assistive environment must be able to perform.

NOTES

1. Monekosso, Remagnino & Kuno, 2009.
2. Jönssen, "Allmänt om rehabiliteringsteknologi."
3. Personal digital assistant, an electronic device which can include some of the functions of a computer, a cellphone, a music player, and a camera.
4. Chan et al., "Review of Smart Homes."
5. Thomas, "Disability Theory," s. 43.
6. Disablism is discriminatory, oppressive, or abusive behavior arising from the belief that disabled people are inferior to others.
7. Barnes and Mercer, "Granskning av den sociala handikappmodellen."
8. Grönvik, "Funktionshinder"; Stone, *Disabled State*.
9. Switsky and Greenspan, *What Is Mental Retardation?*.
10. Helal, Mann, and Lee, "Assistive Environments," 381.
11. Baldoni et al., "Embedded Middleware Platform."
12. Cook, "Prediction Algorithms"; Jönssen, "Allmänt om rehabiliteringsteknologi."
13. Baldoni et al., Embedded Middleware Platform."
14. Abowd and Mynatt, "Designing for the Human Experience."
15. Babbitt et al., "Privacy Management."
16. Helal, Mann, and Lee, "Assistive Environments."
17. Ringsby-Jansson, 2002.
18. Gough and Anderson, 2004.
19. Babbitt et al., "Privacy Management."

BIBLIOGRAPHY

Abowd, Gregory D., and Elizabeth D. Mynatt. "Designing for the Human Experience in Smart Environments." In *Smart Environments: Technology, Protocols, and Applications*, edited by Diane Cook and Sajal Das. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

Babbitt, Ryan, Johnny Wong, Simanta Mitra, and Carl Chang. "Privacy Management in Smart Homes: Design and Analysis." In *Promoting Independence for Older Persons with Disabilities*, edited by Abdelsalam Helal and William C. Mann. Amsterdam: IOS Press, 2006.

Baldoni, R., C. Di Ciccio, M. Mecella, et al. "An Embedded Middleware Platform for Pervasive and Immersive Environments for-All." In 2009 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops. IEEE, 2009.

Barnes, C., and G. Mercer. "Granskning av den sociala handikappmodellen." In *Handikapp: synsätt, principer, perspektiv*, edited by M. Tideman. Studentlitteratur, Lund, 2000.

Chan, Marie, Daniel Estève, Christophe Escriba, and Eric Campo. "A Review of Smart Homes—Present State and Future Challenges." *Computer Methods and Programs in Biomedicine* 91 (2008): 55–81.

Cook, Diane J. "Prediction Algorithms for Smart Environments." In *Smart Environments: Technology, Protocols, and Applications*, edited by Diane Cook and Sajal Das. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

Gough, Ritva and Laila Andersson. *Bostäder med särskild service: en kartläggning av boende för människor med begåvningshandikapp*. Kalmar: Fokus Kalmar lan, 2004.

Grönvik, L. "Funktionshinder – ett mångtydigt begrepp." In *Forskning om funktionshinder, problem-utmaningar-möjligheter*, edited by M. Söder. Studentlitteratur, Lund, 2005.

Helal, Abdelsalam, William C. Mann, and Choonhwa Lee. "Assistive Environments for Individuals with Special Needs." In *Smart Environments: Technology, Protocols, and Applications*, edited by Diane Cook and Sajal Das. Hoboken, NJ: John Wiley & Son, Inc., 2004.

Jönsson, B. "Allmänt om rehabiliteringsteknologi, design och pedagogik." In *Människonära Design*. Studentlitteratur, Lund, 2005.

Ringsby-Jansson, Bibbi. *Vardagslivets arenor: om människor med utvecklingsstörning, deras vardag och sociala liv*. Diss. Göteborg: Univ., 2002.

Stone, Deborah A. *The Disabled State*. Philadelphia: Temple University Press, 1984.

Switsky, Harvey N., and Stephen Greenspan. *What Is Mental Retardation?* Washington, D.C.: AAMR, 2003.

Thomas, Carol. "Disability Theory: Key Ideas, Issues and Thinkers." In *Disability Studies Today*, edited by Colin Barnes, Mike Oliver, and Len Barton. Cambridge: Polity Ltd., 2002.

A Brief Overview of the Philosophical Study of Computer Simulations

Juan M. Durán
UNIVERSIDAD NACIONAL DE CÓRDOBA (ARGENTINA),
JUANDURAN@GMAIL.COM

INTRODUCTION

There is general agreement in the scientific community that computer simulations represent a distinctive method for the sciences. However, it is not sufficiently clear what are the philosophical implications involved in these new methods. One may ask: What does it make a computer simulation different from any other highly successful method in the history of science, such as the use of Pascal's arithmetic machine for calculations, or the employment of the astrolabe for predicting the position of planets? One possible answer is that computer simulations are not specific to a discipline, but they "adapt" to specific scientific needs. For instance, in the physical sciences it is more common to find equation-based computer simulations, whereas in the social sciences agent-based simulations are more suitable. In addition, computer simulations are significantly more powerful than the calculator, and provide more accurate data than the astrolabe. These few highlighted features represent a methodological and epistemic advantage of computer simulations over other scientific instruments. In particular, their adaptability, speed, and accuracy can be cashed out in terms of the kind of knowledge that simulations provide, the kind of theories they contribute to build, confirm, and refute, and the kind of evidence they supply.

Naturally, these issues have attracted the interest of philosophers across disciplines. However, despite the renewed enthusiasm in the past few years for analyzing computer simulations, the philosophical questions that lie behind these practices have been around for some time. As early as 1967, Naylor, Burdick, and Sasser were already discussing definitions of computer simulations for an economic system:[1]

> A numerical technique for conducting experiments with certain types of mathematical and logical models describing the behavior of an economic system on a digital computer over extended periods of time. . . . The principal difference between a simulation experiment and a "real world" experiment is that with simulation the experiment is conducted with a model of the economic system rather than with the actual economic system itself.[2]

It is astonishing to note the similarity of this quotation with more contemporary literature on the topic. Current philosophical inquiry also engages in similar efforts, as it is distinguishing between a computer simulation and a "real world" experiment, or exploring the methodological implications of implementing a scientific model into the computer simulation. Yet, despite these few listed similarities, more needed to be said. From an historical perspective, the introduction of silicon based circuits helped in the standardization of computational architecture and general reliability of the new machines. The growth in speed of calculation, size of memory, or number of programming languages forcefully challenged established ideas and encouraged the seeking of new questions and answers.

In 1990, Fritz Rohrlich conceptualized computer simulations as lying "somewhere intermediate between traditional theoretical physical science and its empirical methods of experimentation and observation."[3] The quotation sets the mood for analyzing computer simulations in the light of a philosophy of models and/or experiments. Arguably, it is still customary to locate the study of computer simulation within these philosophical frameworks. However, as I will show later, there are also successful efforts in addressing the philosophy of computer simulations at face value; that is, by analyzing what is so characteristic that makes them central for the scientific enterprise. In this article, then, I propose to briefly review the most recent philosophical literature on computer simulations. For this, I focus on a selected class of computer simulations, as well as on a selected set of authors. Specifically, I am interested in the literature specialized on equation-based computer simulations (e.g., partial differential equations, ordinary differential equations, and the like). The universe of computer simulations also includes cellular automata, agent-based simulations, and complex systems. However, it is important to discriminate the class of computer simulation of interest since the philosophical analysis is sensitive to what they have to offer. For instance, cellular automata are discrete systems, whereas equation-based simulations implement continuous equations. This is reflected in that a cellular automata would provide exact results of the models implemented whereas equation-based simulations experience all kinds of errors in their results, such as round-off errors, truncation errors, and the like.[4]

That said, let me now identify three moments in the contemporary philosophical literature of computer simulations. The first moment is characterized by attempts to define computer simulations; most notably is the work of Humphreys (1990; 2004) and Hartmann (1996). During the second moment, however, the interest shifts to analyzing similarities (or differences thereof) between computer simulations and laboratory experimentation. The list of authors focused on this second moment is significantly large; a chronological approximation would include Winsberg (1999), Guala (2002), Morgan (2003; 2005), and Morrison (2009), just to mention a few. In this paper, I shall only focus on a few authors. The third moment is characterized by the analysis of computer simulations in their own right; that is, regardless of their similarities or differences with laboratory experimentation, and regardless of approximate definitions. Most notably is the work of Barberousse et al. (2009), Humphreys (2013), and Tal (2009), among others. To my mind, the transition to the third moment has been possible (or at least partly possible) due to the strong criticism that the philosophical studies on computer simulations have received from Frigg and Reiss (2009). But enough of preambles, let me now begin with the first moment.

FIRST MOMENT: THE DEFINITION

The first moment is characterized by a few attempts to define computer simulations. In 1990, Paul Humphreys wrote a stimulating paper where he presents his "working definition" for computer simulations. In it, the author maintains that scientific progress needs, and is driven by, tractable mathematics.[5] Albeit an indubitable point, Humphreys is taking the notion of "mathematical tractability" as excluding cases where the solution of mathematical models is not computable by analytic methods.[6] In other words, "mathematical tractability" here means "to be analytically solvable." Regardless of the differences in opinions that such an interpretation of "tractability" might yield, the important point is that Humphreys considers computer simulations as the new contributors to the history of scientific progress. To his mind, computer simulations "turn analytically intractable problems into ones that are computationally tractable."[7] Computer simulations, therefore, come to amend what analytic methods cannot undertake; that is, to find approximate solutions to equations by means of reliable (and fast) calculation. It is in this context that Humphreys offers the following "working definition":

> *Working Definition*: A computer simulation is any computer-implemented method for exploring the properties of mathematical models where analytic methods are unavailable.[8]

There are two ideas in this working definition worth underlining. The first one has been already discussed; that is, that computer simulations provide solutions to mathematical models where analytic methods are unsuccessful. A follow up comment is that Humphreys is careful in making clear that his working definition should not be identified with numerical methods: whereas both computer simulations and numerical methods are interested in finding approximate solutions to equations, only the latter is related to numerical analysis.[9] The second element is the "exploratory" capacity of computer simulations for finding the set of solutions of the mathematical model. This is certainly a major feature of computer simulations as well as a source for epistemic and methodological discussion. Unfortunately, and to the best of my knowledge, there is little analysis of this topic (a notable exception is García and Velasco, 2013).

Humphreys's working definition was subject of much criticism, especially from Stephan Hartmann (1996). Hartmann objected that one of the main disadvantages is that Humphreys did not take into account the dynamic nature of the implemented model. Instead, he offered an alternative definition:

> Simulations are closely related to dynamic models. More concretely, a simulation results when the equations of the underlying dynamic model are solved. This model is designed to imitate the time-evolution of a real system. To put it another way, a simulation imitates one process by another process. In this definition, the term "process" refers solely to some object or system whose state changes in time. If the simulation is run on a computer, it is called a computer simulation.[10]

Philosophers have warmly welcomed Hartmann's definition. Recently, Wendy Parker has made explicit reference to it: "I characterize a simulation as a time-ordered sequence of states that serves as a representation of some other time-ordered sequence of states."[11] Francesco Guala (2002) also follows Hartmann in distinguishing between *static* and *dynamic* models, time-evolution of a system, and the use of simulations for mathematically solving the implemented model.

In spite of this acceptability, Hartmann's definition presents some problems of its own. The overall assumption is that a computer simulation is the result of direct implementation of a dynamic model on the digital computer, as it follows from the first, second, and last sentences of the previous quotation. To Hartmann's mind, therefore, there is no conceptual difference between solving a dynamic model and running a computer simulation, for the latter is simply the implementation of the former on the digital computer. However, the diversity of methods and processes involved in the implementation of a dynamical model on the digital computer exceed any interpretation of "direct implementation." There is generalized agreement among philosophers on the importance of analyzing the methodology of computer simulations for their conceptualization.[12]

In addition, Hartmann's definition is at odds with taking computer simulations as philosophically interesting and conceptually novel objects of inquiry. More concretely, he seems to consider that the philosophical analysis of computer simulations is subsidiary of the philosophy of models. Indeed, Hartmann defines a computer simulation in terms of a dynamic model; that is: a simulation *results when* the equations in the dynamic model are solved; the dynamic model *is designed to* imitate the time-evolution of a real system; and finally, imitating the time-evolution of a system *is the role ascribed to* the simulation. In this way, a simulation is defined in terms of a dynamic model which, when implemented on the digital computer, is conceived as a simulation. To put the same idea in a slightly different way: a dynamic model (on the computer) becomes a simulation while a simulation is the dynamic model (on the computer). The notion of computer simulation, then, is simplified by defining it in terms of a dynamic model running on a special digital device. It follows that there is nothing particularly special about computer simulations, for any philosophical issue related to the latter can be answered by the philosophy of scientific models.

Despite Hartmann's considerations, there are reasons for thinking that the philosophical study on computer simulations is not a subchapter of the philosophy of models (nor of the philosophy of experimentation, as I argue in the second moment). Of particular interest is Humphreys's reply to Frigg and Reiss, and the way he notices that philosophically motivated questions from the philosophy of models only illuminates one side of the problem; that is, those issues that are also shared by the philosophy of computer simulations. However, such approach obscures the philosophical analysis of computer simulations in itself, such as their methodology, their ontology, and their epistemology. There is, nevertheless, one common perspective shared between Humphreys and Hartmann: both authors agree in considering computer simulations as reckoning devices; that is, high-speed abacus for solving analytically unsolvable mathematical models.[13] Admittedly, calculability and mathematical tractability are

among the most conspicuous features highlighted in current definitions of computer simulation.

After Hartmann's initial objections, Humphreys coined a new definition, this time based on the notion of "computational template."[14] Briefly, a computational template is the result of a computationally tractable theoretical template. A theoretical template, in turn, is the kind of very general mathematical descriptions that can be found in a scientific work. This includes partial differential equations, such as elliptic (e.g., Laplace's equation), parabolic (e.g., the diffusion equation), and hyperbolic (e.g., the wave equation), ordinary differential equations, among others.[15] An illuminating example of a theoretical template is Newton's Second Law: "[it] describes a very general constraint on the relationship between any force, mass, and acceleration."[16] These theoretical templates need to be specified in some particular respects, for instance, in the force function: it could either be a gravitational force, an electrostatic force, a magnetic force, or any other variety of force. Finally, "if the resulting, more specific, equation form is computationally tractable, then we have arrived at a computational template."[17] Arguably, this is less a definition as it is a characterization of the notion of computer simulation. In any case, it is, together with Hartmann's, the most popular conceptualization among current philosophical literature.

Let me finish this section with an illuminating classification of the term "computer simulation" as construed by Roman Frigg and Julian Reiss. According to the authors, there are two senses in which the notion of "computer simulation" is defined in current literature:

> In the *narrow* sense, "simulation" refers to the use of a computer to solve an equation that we cannot solve analytically, or more generally to explore mathematical properties of equations where analytical methods fail (e.g., (Humphreys 1991, p. 501; 2004, p. 49; Winsberg 1999, p. 275; 2001, p. 444).
>
> In the *broad* sense, "simulation" refers to the entire process of constructing, using, and justifying a model that involves analytically intractable mathematics (e.g., Winsberg 2001, p. 443; 2003, p. 105; Humphreys 1991, p. 501; 2004, p. 107). Following Humphreys (2004, pp. 102–104), we call such a model a "computational model."[18]

Both categories are certainly meritorious and illuminating. Both capture the two senses in which philosophers define the notion of "computer simulation." While the narrow sense focuses on the heuristic capacity of computer simulations, the broad sense emphasizes the methodological, epistemological, and pragmatic aspects of computer simulations.

After this initial momentum, the efforts in understanding computer simulations shifted to the comparison with laboratory experimentation and scientific models (both broadly construed). Let me now address this shift in the "second moment."

SECOND MOMENT: THE COMPARISON

Perhaps the most pressing philosophical question about computer simulations comes from asking about their epistemic power. Philosophers, encouraged by actual scientific practice, tend to address this issue by comparing computer simulations *vis à vis* experimental practices. This is the distinctive hallmark of the second moment; that is, to illuminate the study of computer simulations by means of comparing them with laboratory experimentation.

Now, comparisons are never as straightforward as they seem. One of the major problems philosophers run into was that the philosophy of science is markedly empirical, with observation and experimentation of real-world phenomena at its center. Computer simulations, on the other hand, are entrenched as being neither purely empirical nor entirely theoretical, but rather "[lying] somewhere intermediate between traditional theoretical physical science and its empirical methods of experimentation and observation."[19] The "first moment" taught us that many computer simulations have as target system real-world phenomena, although only in a representative form via the implementation of a mathematical model. The question that rises is, therefore, how can a comparison of the epistemic power be carried out when computer simulations are markedly abstract and experiments are distinctively empirical. Most philosophers have chosen to answer this question by overthrowing the (allegedly) ontological gap existing between computer simulations and experiments, for in this way they can illuminate their epistemic resemblances. Let me now show how this has been formulated in the literature.

The most celebrated criterion for analyzing computer simulations and laboratory experiments is the so-called *materiality argument*. Parker has made the most comprehensible reconstruction of this argument in the following way:

> In genuine experiments, the same "material" causes are at work in the experimental and target systems, while in simulations there is merely formal correspondence between the simulating and target systems . . . inferences about target systems are more justified when experimental and target systems are made of the "same stuff" than when they are made of different materials (as is the case in computer experiments).[20]

A simple analysis shows that the first sentence considers computer simulations as abstract entities, whereas experiments share the same "material" causes with real-world phenomena.[21] The second sentence, however, is epistemic in nature, and aims to assert that sharing the "same stuff" justifies better our epistemic inferences about the target system. It follows that being "material" justifies better (than being abstract) our inferences about the real-world phenomenon. Admittedly, not every philosopher shares this last claim; many also take computer simulations as genuine experiments and, as such, see them to be epistemically on a par with experiments. To my mind, there are three general philosophical stances on this issue; namely:

(a) Computer simulations and experiments are causally similar; hence, both are epistemically on a par. For instance Parker (2009);

(b) Computer simulations and experiments are materially dissimilar, whereas the former is abstract in nature, the latter shares causal relations with the phenomenon under study; hence, both are epistemically different. For instance Guala (2002), Giere (2009), and Morgan (2003, 2005);

(c) Computer simulations and experiment are ontologically similar, i.e., both are model-shaped; hence, both are epistemically on a par. For instance Morrison (2009) and Winsberg (2009).

The important point to note here is the rationale of these arguments. Whether the epistemic payoff is positive or negative, these philosophers consider that the epistemic evaluation of computer simulations find support on a previous ontological commitment. I argued elsewhere (Durán, 2013) that there is a common rationale guiding these different positions, all of which share the logic of the "materiality argument" as common source. I referred to this rationale as the *materiality principle*, and I characterized it as determining the epistemic power of computer simulations (and experiments) by adopting a previous ontological commitment. Putting the same idea in a slightly different form, I consider that there is an ontological commitment to the model abstractness of computer simulations (or to the worldliness/abstractness of experimentation) that determines their respective epistemic power. One of the aims I have for this section is to spot this rationale in all of the three stances aforementioned.

Let me begin the discussion on a non-chronological way by first addressing the work of Wendy Parker (2009). Following Hartman's definition, Parker first draws a distinction between *computer simulations* and *computer simulation studies*. To her mind, a computer simulation is an abstract entity lacking of the typical intervening mechanism present in experimental practice; for this reason, computer simulations do not qualify as experiments. Instead, she coins the notion of *computer simulation study* as a way to include the alleged intervening mechanism. In this way, she equates methodologically and ontologically computer simulations to experiments. A *computer simulation study*, therefore, consists of

> The broader activity that includes setting the state of the digital computer from which a simulation will evolve, triggering that evolution by starting the computer program that generates the simulation, and then collecting information regarding how various properties of the computing system, such as the values stored in various locations in its memory or the colors displayed on its monitor, evolve in light of the earlier intervention (i.e., the intervention that involves setting the initial state of the computing system and triggering its subsequent evolution).[22]

So defined, a computer simulation study does qualify as an experiment insofar it includes the more extensive and human related activity of setting up the simulation. According to Parker, then, there is only one purpose for setting the states of the computer simulation, and that is to collect information on the evolution of the system. Now, such information comes in the form of values stored in the memory, the colors displayed in the monitor, and similar physical states of the digital computer. Interpreted in this way, it seems that Parker also considers that a computer simulation study is the intervening material that *brings about* the phenomenon. This last point is supported by her own words: "I want to stress the importance of instead understanding computer experiments as, first and foremost, experiments on real material systems. The experimental system in a computer experiment is the programmed digital computer (a physical system made of wire, plastic, etc.)."[23]

The motivation that supports her claim is the need to find an answer to the materiality argument. Indeed, in her search for analyzing the epistemic value of computer simulations, Parker attempts to ontologically equate computer simulations studies with laboratory experimentation. This attempt is, in turn, motivated by the analysis of computer simulations in the context of the philosophy of experimentation. The mark of the second moment can be clearly distinguished now: Parker puts computer simulations and experimentation on the same ontological footing, making the former "look like" the latter, for only then she can engineer an evaluation of the epistemic power of computer simulations.

Let me now sketch a possible objection to her viewpoint. In the previous quotation, Parker refers to "collecting information of properties of the computer system" as a way of arguing for the possibility of accessing the various locations in the computer memory, since it is in this way in which the scientist can observe the evolution of the system after a previous intervention.[24] If this is the correct interpretation, then her claim is false. It is impossible for any person to access the different locations of the computer memory (computer bus, processes states, etc.) with the purpose of understanding them. In 2004 Humphreys argued that computer systems hold a certain degree of *epistemic opacity* insofar they are inaccessible for direct inspection and verification.[25] I believe this claim applies to this case and represents a serious objection to Parker's views. Moreover, I have defended a modified multi-realizability argument for computer simulations that applies specifically to Parker's case.[26] In there I vindicate the idea that even in the hypothetical case that the scientist can access the internal states of the computer system, there is no possibility of real understanding since those states are not representative of the states of the real-world phenomenon.

Another interesting effort in differentiating computer simulations from laboratory experimentation is carried out by Francesco Guala. In his 2002 paper, Guala considers that the fundamental difference between computer simulations and experiments lies in whether the "same material causes" are at work in the experimental/simulating and target system. The author explains the differences in the following way:

> The difference lies in the kind of relationship existing between, on the one hand, an experimental and its target system, and, on the other, a simulating and its target system. In the former case, the correspondence holds at a "deep," "material" level, whereas in the latter the similarity is admittedly only "abstract" and "formal." . . . In a genuine experiment the same 'material' causes as those in the target system are at work; in a simulation they are not,

> and the correspondence relation (of similarity or analogy) is purely formal in character.[27]

Same and *different* are the notions that the author uses for emphasizing changes of materiality. The case of the ripple-tank is paradigmatic in this sense. According to Guala, the media in which waves travel are made of different "stuff": while one media is water, the other is light. The ripple-tank, however, can be used as a representation of the wave nature of light only because there are similarities in the behavior of water shared at a very abstract level only (i.e., at the level of Maxwell's equations, D'Alambert's wave equation, and Hook's law). The two systems (water and light), then, obey the same laws and can be represented by the *same* set of equations, despite their being made of *different* "stuff." To Guala's mind, however, no abstract similarity can compensate the difference in materiality: water waves are not light waves, and there is no reductionist story that can bring these two phenomena on equal footing.[28] On the epistemic side, any differences in the materiality presuppose a difference in our understanding and knowledge of the phenomenon "light."

By an analogous reasoning, Guala considers computer simulations and experiments as being materially dissimilar and, therefore, epistemically different. In particular, Guala considers that experiments remains the most reliable way to know something about the real world, downplaying in this way the role of computer simulations in the scientific enterprise. Computer simulations, however, become an attractive method when scientists find difficulties in carrying out controlled experimentation. The example used by the author is a computer simulation used by geologists for stratigraphy.[29] Because the study of sedimentary and layered volcanic rocks is expensive, time-consuming, inaccessible (time, place, dimension), or any of the many reasons that prevent scientists from carrying out direct experimentation, computer simulations make it to central stage in scientific practice. I believe that this tendency towards a disjunctive assessment of experimentation and simulation is a natural consequence of adopting the *materiality principle*.

In a similar fashion, Mary Morgan follows Guala in considering experiments as epistemically privileged over computer simulations. In a couple of very insightful papers, Morgan presents the most exhaustive and rich analysis in current literature. Her main concern is the so-called *vicarious experiments*; that is, "experiments that involve elements of nonmateriality either in their objects or in their interventions and that arise from combining the use of models and experiments, a combination that has created a number of interesting hybrid forms."[30] The importance of studying vicarious experiments is that their epistemic validity is a function of their materiality. More specifically, a vicarious experiment is characterized by what Morgan calls the "degrees of materiality" of experiments; that is, the different degrees by which the materiality of an object is present in the experimental setup (see table ii on page 230, where also includes the two extreme cases: "Ideal Laboratory Experiments" and "Mathematical Model Experiments"). The general epistemological analysis, then, is a function of the degree of materiality of the kind of experiment. In more familiar parlance, back inference to the world is better justified when the experiment and the target system are made of the same "material." In Morgan's own words, "ontological equivalence provides epistemological power" and "the ontology matters because it affects the power of inference."[31]

The ratio "materiality-epistemic power" measures the expected epistemic insight of using a computer simulation or an experiment. For instance, since a "mathematical model experiment" can only represent things in the world, they cannot be used for confirmation of theories. In a similar vein, computer simulations cannot test theoretical assumptions of their target system because it has been designed for delivering results consistent with built-in assumptions. A laboratory experiment, on the other hand, has been explicitly designed for letting the facts about the target system "talk" by themselves, and therefore they are more reliable in terms of information about the world.

It is precisely the material substratum underlying an experiment the responsible for their epistemic power. Morgan draws a table where she shows the limitations of each activity depending on their degree of materiality.[32] For instance, the "ideal laboratory experiment" is epistemically more powerful than a "virtually experiment"; in turn, a "virtually experiment" is more powerful than a "virtual experiment," and so on. Since a computer simulation can only be conceived as a "hybrid experiment" or as a "mathematical experiment," it follows that it is epistemically less powerful than the "ideal laboratory experiment," which is the exemplary experimental case. In other words, the ratio "materiality-epistemic power" can be succinctly put as "the degree of materiality" determines "the degree of epistemic power," which is consistent with the materiality principle.

It is in this context of "degrees of materiality" where Morgan coins the terms "surprise" and "confound." These terms represent the epistemic states in which the scientist enters when presented with results of a computer simulation or results of a material experiment, respectively. Results of a computer simulation can only "surprise" the scientist for its behavior can be traced back to, and explained in terms of, the underlying model. A material experiment, on the other hand, can "surprise" as well as "confound" the scientist, for it can bring up new and unexpected patterns of behavior inexplicable from the point of view of current theory.[33] The materiality of the experiment, then, works as the epistemic guarantee that the results may be novel, as opposed to the simulation, which takes results explainable in terms of the underlying model.

Unlike the last two previous accounts, where computer simulations were epistemically downplayed, Margaret Morrison considers that there are good reasons for raising computer simulations to the level of genuine experiments. In this sense, Morrison shares the sentiment with Parker for engineering computer simulations as experimental devices (let us remember that this move supports the epistemic importance of simulations). However, unlike every argument I have discussed so far, Morrison astutely shifts the burden of proof in the ontological analysis from computer simulations to analyzing experiments. Let me explain this point a bit further: most philosophers take for granted that experiments are "material" in a straightforward sense, whereas computer simulations are abstract and formal. The philosophical analysis, then, turns to shape differences and

similarities that would endorse their epistemic views. For those philosophers that support computer simulations as experiments, the philosophical analysis turns to quantify the former as ontologically similar to the latter (e.g., Parker, as I discussed previously). Morrison, instead, bolsters the ontological analysis of experiments as opposed to computer simulations. In this vein, she takes experiments as model-shaped in a straightforward sense, yielding the conclusion that computer simulations are, indeed, experimental devices. The philosophical challenge, then, consists in showing why experiments are model-shaped and, as such, closer to computer simulations. The general strategy of her 2009 paper is to show that certain types of computer simulations share the same ontological characteristics as experimental measurements, and therefore both epistemically on a par.

Allow me now to illustrate her account by briefly addressing the example of measuring *g*.[34] In an experimental measurement, Morrison argues, a scientific instrument measures a physical property up to certain degree of *precision*, although such measurement will not necessarily reflect an *accurate* value of the property in question. Now, the distinction between *precision* and *accuracy* is of paramount importance for Morrison; whereas the former is related to the experimental practice of intervening in nature, the latter is related to the mediation of models as render of reliable data. In this context, a *precise measurement* consists of the set of results whose degree of uncertainty in the estimated value of a physical property is relatively small;[35] on the other hand, an *accurate measurement* consists of the set of corrected results that are close to the true value of the measured property.[36] The difference between these two concepts is the cornerstone to Morrison's strategy. Data collected from experimental instruments only provide *precise* measurements of *g*, whereas reliable measurements must be primarily accurate representations of the value measured, and therefore post-processed in the search for such *accuracy* (for the particular case of measuring *g*, Morrison proposes the ideal point pendulum as theoretical model).

From Morrison's perspective, then, the reliability of the measured data is a function of the level of accuracy, which depends, in turn, on a theoretical model. In her own words:

> The calculation generates a large amount of data which requires that they be appropriately modelled in order to render them interpretable. Only by doing that can we say that the computer experiment, like an ordinary experiment, has measured a particular quantity. In both cases models are crucial. And, just as in the pendulum example where we are interested in both the precision and accuracy, similar concerns arise for simulation where the precision of the machine and the behaviour of apparatus is related to the observed properties of the microscopic system.[37]

Scientific instruments and computer simulations share the same fate of being precise but not accurate. The former, due to the physical constraints related to manipulating and intervening the real world (e.g. a pendulum measuring the gravitational force of the Earth). The latter, because the computer simulation includes, in its model, the physical constraints of the target system as well as the physical constraints of the machine itself (e.g., round-off errors, truncation errors, and so forth). The dichotomy precision/accuracy, then, equally applies to computer simulations as it does for experimental measurement, making both practices ontologically equal at the level of precise data, and epistemically equal at the level of accurate data.

In this simple way, Morrison ontologically identifies models and computer simulations, which incidentally fulfill the same epistemic role: "by focusing on how models function as measuring instruments in experimental inquiry I have tried to shed some light on the way simulations, as a form of modelling, can fulfill the same role."[38]

Morrison's paper faced some resistance from philosophers that expect to keep the empirical aspects of experimentation intact. Ronald Giere (2009), for instance, wrote a direct response to Morrison's paper claiming that computer simulations lack of any causal interaction with real phenomena, and therefore they cannot confer any additional confidence in the fit between the simulation model and the target system. To Giere's mind, a computer simulation cannot go beyond mere calculation, an abstract and (sometimes) formal process *par excellence*. Laboratory experiment, on the other hand, holds the distinctive characteristic of interacting and manipulating real-world phenomena through physical causal relations. To his mind, then, there is an ongoing confusion about the nature of each activity. Following Winsberg (2009), Giere considers that the epistemic role of computer simulations is to make predictions, calculate complex equations, and eventually to agree with empirical data, but above and beyond those few mechanistic processes, computer simulations cannot overthrow experimental practice. As Giere explains it,

> Our epistemological confidence in the ability of the simulation model adequately to represent the target system rests on our confidence in the fundamental principles built into the simulation model plus the known reliability of various modeling techniques. Computer experiments not connected to actual data do not confer any additional confidence in the fit (or lack thereof) between the simulation model and the target system.[39]

Giere's conclusion is that computer simulations are materially different from experiments and, as such, epistemically less powerful. In this sense, Giere's account is closer to Parker's, and Morgan's.

I must now desist from carry out any more analysis. The list of philosophers dedicated to the study of computer simulations is surprisingly long, including the numerous works of Winsberg[40] (1999, 2003, 2009), Lenhard (2007), Norton and Suppe (2001), Hughes (1999), Arnold (2013), among others. To my mind, these authors conceive the study of computer simulations mainly as subsidiary to the philosophy of experimentation (or models, as I pointed out in the first moment); that is, the philosophy of computer simulations is illuminated by the more familiar problems of the philosophy of experimentation.[41]

Such unidirectional approach highlights only the similarities (and differences) between computer simulations and

experiments, but it never shows what is interesting about computer simulations in and by themselves. For instance, one can argue that computer simulations cannot act as *experiment crucis* (Arnold, 2013), without saying much about the reliability of computer simulations in procedures for detecting phenomena (Tal, 2011). The third moment is precisely marked by the philosophical analysis of computer simulation in and by themselves. The lesson of this third moment will be that the study of computer simulations emerges as a discipline of genuine philosophical interest.

THIRD MOMENT: HOT NEW STEW

In 2008, Roman Frigg and Julian Reiss wrote an insightful paper questioning the importance of analyzing computer simulations as a separate topic in the philosophy of science. To their mind, there is an overstated interest in computer simulations, followed by bogus demands for a new philosophy of science. I am skeptical to say that Galison (1996), Winsberg (1999, 2001), Humphreys (2004), or Rohrlich (1991), the four authors mentioned in the paper, are demanding for a new philosophy of science. Rather, I understand their claim as a rejection to considerations that computer simulations are a subchapter of a more familiar philosophy (either a philosophy of models, or a philosophy of experiment), precisely as Frigg and Reiss suggest. Alternatively, these philosophers claim that the philosophical interest in computer simulations lies on the ontological shift simulations impose with respect to more traditional scientific practice (Galison, 1996); the complex chain of inferences that transform theoretical structures into concrete knowledge of physical systems (Winsberg, 1999); or simply the ubiquitousness of computer simulations in scientific practice. Nothing of which, I believe, suggests that the past eighty or ninety years of philosophy of science must be rewritten.

Nevertheless, it is always a healthy intellectual exercise to raise doubts about the importance of philosophically studying new methods in the sciences. For this, Frigg and Reiss based their case on four claims; namely,

> *Metaphysical*: Simulations create some kind of parallel world in which experiments can be conducted under more favourable conditions than in the "real world."
>
> *Epistemic*: Simulations demand a new epistemology.
>
> *Semantic*: Simulations demand a new analysis of how models/theories relate to concrete phenomena.
>
> *Methodological*: Simulating is a Sui Generis activity that lies "in between" theorising and experimentation.[42]

According to the authors, then, philosophers of computer simulations have construed and based their arguments on these four claims, none of which supports anything remotely close to a "new epistemology." In fact, Frigg and Reiss assert that computer simulations "raise few if any new philosophical problems," and that any issue related to computer simulations is of another order, whereas mathematical, physical, or even psychological, but definitely not philosophical.[43]

At this point is where opinions tend to diverge. Humphreys, for instance, wrote an insightful response to Frigg and Reiss's paper. In there, he addresses each one of the aforementioned claims defending computer simulations as a novel scientific practice that raises genuine philosophical interest. In particular, he elaborates on what he called the *anthropocentric predicament*, which asserts in the following question: "how we, as humans, can understand and evaluate computationally based scientific methods that transcend our own abilities?"[44] The anthropocentric predicament is designed to question the established empiricist-based philosophy of science, whose center is humans and their capacities to observe and experiment. Instead, computer simulations use "methods that push humans away from the centre of the epistemological enterprise," making the philosophical study of computer simulations an important enterprise in itself.[45]

I consider Humphreys's paper to mark the transition into what I called the "third moment": philosophers primarily focus on features and characteristics of computer simulations in themselves, instead of on a comparing basis with laboratory experimentation or models. In this respect, the work of Barberousse, Franceschelli, and Imbert (2009) on empirical and computer data is a good example of this;[46] Humphreys has also done some work on the content of data produced by a computer simulation. Allow me now to briefly address these views.

Born into the discussion on whether computer simulations are comparable to material experiments, Barberousse et al. turn their attention to the data produced by a computer simulation. The importance of their work lies in the fact that it provides a qualified answer to the question about the type of data that computer simulations produce, and whether such data resemble in any way data produced by an experiment. Strictly speaking, the authors are still engaged in the discussion about the physicality of the digital computer and its role in the epistemology of computer simulations. However, instead of comparing computer simulations *vis à vis* laboratory experiments from the point of view of the philosophy of experiment, the authors focus on specific characteristics of computer simulations, imprinting in this way a shift in the philosophical analysis.

The motivation behind Barberousse's paper comes from Humphreys's 1994 article, where he calls for attention to the semantic problem of "numerical experimentation." Briefly, Humphreys notices that the Ising model has no analytic solution for three-dimensional lattices, resulting in analytic intractable integrals. The Monte Carlo method, then, is presented as the best approximation to the solution of these integrals, in addition to be entirely devoid of empirical content: "none of the inputs come from measurements or observations on real systems, and the lattice models are just that, mathematical models."[47] Now, although Humphreys was concerned about the mathematical use of computers for solving intractable equations (method that, according to the author, does not belong neither to physical experimentation nor numerical mathematics), Barberousse et al. prefer to focus on the semantic content that the production of data presupposes. In this sense, Barberousse et al. divide data into two kinds: $data_E$ and $data_A$. The former is data produced by physical interactions with measuring or detecting devices,

as it can be a pendulum measuring the gravitational force *g*; the latter is considered as data about a physical system, as 9.8 m/s^2 is data about the force *g*. These two types of data are not necessary disjoint, for data about a system ($data_A$) may be produced by data from an empirical origin ($data_E$), as it is the case of data obtained from using a pendulum is also data about the gravitational force *g*. In addition, $data_A$ can also be obtained from pen-and-pencil calculations, as it is the case of calculating systems of equations.[48] After an insightful analysis of computer simulation data, the authors conclude that the dividing line between simulations and laboratory experiments can be drawn using precisely their semantic analysis.

Barberousse's semantic analysis focuses primarily on two elements; that is, on the source of production of data, and on the representational capacity of the target system's properties. According to Humphreys, however, the origins of data about ($data_A$) are insufficient to determine the content of such data: "decisions about $data_A$ require knowledge of what causal processes were involved in producing the individual data and what transformations have been performed on the individual data points."[49] His argument gets its force from what the author calls "causal-computational instruments"; that is, instruments that take physical processes as inputs and, at some point, they convert these physical states into digital representation suitable for undergoing computational transformations. The decisive point is that the data delivered by a causal-computational instrument are the result of deliberate engineering. Depending on the particular purpose, say whether the data is meant to be "read" by a human agent, or further processed in the computer, the appearance of the engineered data may differ greatly. In order to determine its representational content, it is therefore crucial to consider the origins of the data as well as the engineering steps by which it is formed (and transformed). Despite the seemingly arbitrary process of constructing data about a target system, Humphreys indicates, "the output will be tailored to the needs of the data user, whether it is a human scientist, an automated scientist, or some other epistemic agent."[50]

One genuine question to ask is what is the connection between "causal-computational instruments" and computer simulations. Humphreys indicates that many of his claims can be transferred to the case of computer simulations. The problem now lies on how to interpret $data_A$, for it can be determined by reference to the simulation data itself, or by an intentional attribution to the output from the simulation. In either way, computer simulations (in their form of causal-computational instruments) pose a significant challenge for philosophers interested in philosophical problems of realism, data, and similar topics.

In a way, the interest of Barberousse et al. and Humphreys for comparing computer simulations and laboratory experiments has not ceased, and there are no reasons for this to happen. The interesting aspect about the third moment is, however, the shift towards a more directed philosophy of computer simulations, rather than an elliptical approach through analyzing models or laboratory experiments. This shift allows focusing on features that computer simulations have to offer by themselves, and how their analysis contributes to the philosophy of science broadly construed.

ACKNOWLEDGEMENTS

I am grateful to Pío García and Marisa Velasco (Universidad Nacional de Córdoba, Argentina) for comments on an earlier version of the paper. I am also in debt with Ulrike Pompe (University of Stuttgart, Germany) for a careful reading of this work. Finally, my thanks go to Manuel Barrantes (University of Virginia, USA) for his comments and encouragement.

NOTES

1. One can, of course, find earlier examples. See, for instance, Aspray, *John von Neumann and the Origins of Modern Computing*.
2. Naylor, Burdick, and Sasser, "Computer Simulation Experiments with Economic Systems," 1316.
3. Rohrlich, "Computer Simulation in the Physical Sciences," 507.
4. Toffoli, "Cellular Automata."
5. Humphreys, "Computer Simulation," 497–98.
6. Ibid., 500.
7. Ibid., 501.
8. Ibid.
9. Ibid., 502.
10. Hartmann, "World as a Process," 83.
11. Parker, "Does Matter Really Matter?," 486.
12. Winsberg, "Simulated Experiments"; Humphreys, *Extending Ourselves*; Morgan, "Experiments versus Models."
13. Humphreys, "Computer Simulation," 499–500; Hartmann, "World as a Process," 83.
14. Humphreys, *Extending Ourselves*, 60ff.
15. Ibid., 68.
16. Ibid., 60.
17. Ibid., 60–61.
18. Frigg and Reiss, "Philosophy of Stimulation," 596.
19. Rohrlich, "Computer Stimulation in the Physical Sciences."
20. Parker, "Does Matter Really Matter?," 484.
21. Some of the terminology in the literature remains unspecified, such as "material" causes or "stuff" (Guala, 2002). I take them here to mean "physical causal relations," as described, for instance, by Dowe (2000). In the same vein, when I refer to "causes," "causality," or similar terms, it should be interpreted in the way here specified.
22. Parker, "Does Matter Really Matter?," 488.
23. Ibid., 488–89.
24. Ibid., 488.
25. Humphreys, *Extending Ourselves*, 147.
26. Durán, "Use of the 'Materiality Argument.'"
27. Guala, "Models, Simulations, and Experiments," 66–67.
28. Ibid., 66.
29. Ibid., 68.
30. Morgan, "Experiments without Material Intervention," 217.
31. Morgan, "Experiments versus Models," 324–26.
32. Morgan, "Experiments without Material Intervention," 230.
33. Morgan, "Experiments versus Models," 325; Morgan, "Experiments versus Models," 219.
34. Morrison also discusses the more sophisticated example of spin measurement in "Models, Measurement, and Computer Simulation," 51.
35. Ibid., 49.
36. The difference between precision and accuracy is also explained by Franklin in the following example: "a measurement of the speed of light, c = (2.000000000 ± 0.000000001) x 1010 cm/s is precise but inaccurate, while a measurement c = (3.0 ± 0.1) x 1010 cm/s is

more accurate but has a lower precision." Franklin, "What Makes a 'Good' Experiment?," 367, note 1.

37. Morrison, "Models, Measurement, and Computer Simulation," 53.

38. Ibid., 55.

39. Giere, "Changing the Face of Experimentation," 61–62.

40. Admittedly, it is controversial to include Winsberg among these philosophers. He is aware of the differences between addressing computer simulations at face value and by means of a more familiar philosophy. In this regard, Winsberg says: "Computer simulations have a distinct epistemology. . . . In other words, the techniques that simulationists use to attempt to justify simulation are unlike anything that usually passes for epistemology in the philosophy of science literature" (Winsberg, "Simulations, Models, and Theories," 447). However, I believe he belongs to this section since his work mainly focuses on comparing computer simulations with scientific experimentation.

41. Marisa Velasco pointed out to me two more issues characteristic of this moment—that is, an oversimplification in the philosophical analysis of scientific practice, as well as in the number of computer simulations under investigation. I agree with her in the first claim, but I still find reasons for narrowing down the class of computer simulations to a representative set (say, all equation-based computer simulations). The reason for this has been given at the introduction of this paper; namely, that there are different epistemic properties in different classes of computer simulations. For instance, cellular automata and agent-base simulations have emergent properties that an equation-based simulation not necessarily possesses.

42. Frigg and Reiss, "Philosophy of Simulation."

43. Ibid.

44. Humphreys, "The Philosophical Novelty of Computer Simulations," 616.

45. Ibid.

46. See also Barberousse and Vorms, "Computer Simulations and Empirical Data."

47. Humphreys, "Numerical Experimentation," 112.

48. Barberousse et al., "Computer Simulations as Experiments," 560.

49. Humphreys, "What are Data About?"

50. Ibid.

BIBLIOGRAPHY

Arnold, Eckhart. "Experiments and Simulations: Do They Fuse?" In *Computer Simulations and the Changing Face of Scientific Experimentation*, edited by Juan M. Durán and Eckhart Arnold. Cambridge Scholars Publishing, 2013.

Aspray, William. *John von Neumann and the Origins of Modern Computing*. The MIT Press, 1990.

Barberousse, Anouk, Sara Franceschelli, and Cyrille Imbert. "Computer Simulations as Experiments." *Synthese*, 169, no. 3 (2009): 557–74.

Barberousse, Anouk, and M. Vorms. "Computer Simulations and Empirical Data." In *Computer Simulations and the Changing Face of Scientific Experimentation*, edited by Juan M. Durán and Eckhart Arnold. Cambridge Scholars Publishing, 2013.

Dowe, Phil. *Physical Causation*. New York: Cambridge University Press, 2000.

Durán, Juan M. "The Use of the 'Materiality Argument' in the Literature on Computer Simulations." In *Computer Simulations and the Changing Face of Scientific Experimentation*, edited by J. M. Durán and E. Arnold. Cambridge Scholars Publishing, 2013.

Franklin, Allan D. "What Makes a 'Good' Experiment?" *The British Journal for the Philosophy of Science*, 32, no. 4 (1981): 367.

Frigg, Roman and Julian Reiss. "The Philosophy of Simulation: Hot New Issues or Same Old Stew?" *Synthese* 169, no. 3 (2009): 593–613. doi:10.1007/s11229-008-9438-z.

Galison, Peter. "Computer Simulation and the Trading Zone." In *Disunity of Science: Boundaries, Contexts, and Power*, edited by Peter Galison and David J. Stump. Stanford University Press, 1996.

García, Pio. and Marisa Velasco. "Exploratory Strategies: Experiments and Simulations." In *Computer Simulations and the Changing Face of Scientific Experimentation*, edited by Juan M. Durán and Eckhart Arnold. Cambridge Scholars Publishing, 2013.

Giere, Ronald N. "Is Computer Simulation Changing the Face of Experimentation?" *Philosophical Studies*, 143, no. 1 (2009): 59–62.

Guala, Francesco. "Models, Simulations, and Experiments." In *Model-Based Reasoning: Science, Technology, Values*, edited by Lorenzo Magnani and Nancy J. Nersessian 59–74. Kluwer, 2002.

Hartmann, Stephan. "The World as a Process: Simulations in the Natural and Social Sciences." In *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, edited by Rainer Hegselmann, Ulrich Mueller and Klaus G. Troitzsch, 77–100. Kluwer, 1996.

Hughes, R. I. G. "The Ising Model, Computer Simulation, and Universal Physics." In *Models as Mediators. Perspectives on Natural and Social Science*, edited by Mary S. Morgan and Margaret Morrison, 97–145. Cambridge University Press, 1999.

Humphreys, Paul. "What are Data About?" In *Computer Simulations and the Changing Face of Scientific Experimentation*, edited by Juan M. Durán and Eckhart Arnold. Cambridge Scholars Publishing, 2013.

Humphreys, Paul W. "Computer Simulation." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 2 (1990): 497–506.

———. "Numerical Experimentation." In *Patrick Suppes: Scientific Philosopher Vol. 2*, edited by Paul W. Humphreys, 103–121. Kluwer, 1994.

———. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press, 2004.

———. "The Philosophical Novelty of Computer Simulations." *Synthese* 169 (2009): 615–26.

Lenhard, Johannes. "Computer Simulation: The Cooperation between Experimenting and Modeling." *Philosophy of Science*, 74 (2007): 176–94.

Morgan, Mary S. "Experiments without Material Intervention." In *The Philosophy of Scientific Experimentation*, edited by Hans Radder, 216–235. University of Pittsburgh Press, 2003.

———. "Experiments versus Models: New Phenomena, Inference and Surprise." *Journal of Economic Methodology* 12, no. 2 (2005): 317–29.

Morrison, Margaret. "Models, Measurement, and Computer Simulation: The Changing Face of Experimentation." *Philosophical Studies* 143, no. 1 (2009): 33–57.

Naylor, Thomas H., Donald S. Burdick, and W. Earl Sasser. "Computer Simulation Experiments with Economic Systems: The Problem of Experimental Design." *Journal of the American Statistical Association* 62, no. 320 (1967): 1315–37.

Norton, Stephen D. and Frederick Suppe. "Why Atmospheric Modeling is Good Science." In *Changing the Atmosphere: Expert Knowledge and Environmental Governance*, edited by Clark A. Miller and Paul N. Edwards, 67–105. MIT Press, 2001.

Parker, Wendy S. "Does Matter Really Matter? Computer Simulations, Experiments, and Materiality." *Synthese* 169, no. 3 (2009): 483–96.

Rohrlich, Fritz. "Computer Simulation in the Physical Sciences." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (1990): 507–18.

Tal, Eran. "From Data to Phenomena and Back Again: Computer-Simulated Signatures." *Synthese* 182 (2009): 117–29.

Toffoli, Tommaso. "Cellular Automata as an Alternative to (rather than an approximation of) Differential Equations in Modeling Physics." *Physica D: Nonlinear Phenomena* 10, no. 1–2 (1984): 117–27.

Winsberg, Eric. "Sanctioning Models: The Epistemology of Simulation." *Science in Context* 12 (1999): 275–92.

———. "Simulations, Models, and Theories: Complex Physical Systems and Their Representations". *Philosophy of Science* 68, no. 3 (2001): 442–54.

———. "Simulated Experiments: Methodology for a Virtual World." *Philosophy of Science* 70 (2003): 105–25.

———. "A Tale of Two Methods." *Synthese* 169, no. 3 (2009): 575–92.