NEWSLETTER | The American Philosophical Association

Philosophy and Computers

FALL 2019

VOLUME 19 | NUMBER 1

ISSN 2155-9708

APA NEWSLETTER ON

Philosophy and Computers

PETER BOLTUC, EDITOR

VOLUME 19 | NUMBER 1 | FALL 2019

FROM THE CHAIR

Peter Boltuc

UNIVERSITY OF ILLINOIS SPRINGFIELD

The gist of my term as the chair is dictated by the announcement "Changes to APA Committees" posted on February 23, 2018, especially by the following passage: "the board has made the decision to wind down the Committee on Philosophy and Computers, Committee on Philosophy and Medicine, and Committee on Philosophy and Law."[1] The above webpage allows APA members to post comments, but there were no comments from (or on behalf of) the other affected committees and only one comment at all, by Fritz J. McDonald, at the time a member of this committee, dated February 26, 2018. The gist of Fritz's posting can be summarized by its last sentence: "While information technology grows more and more central to our lives, the APA decides to eliminate the committee on this area." The fact that there was no follow-up does not necessarily mean that nobody agreed with the important aspects of McDonald's comments, but it indicates something more practically important—that there was no overwhelming interest in reopening this issue at the time. I understand the issue to be closed.

Two things would be important to take up. First, the causes and reasons for elimination of this committee. Second, practical issues that pertain to its activities, which include organization of sessions at the APA meetings on philosophical questions in information technology, legacy of the *APA Newsletter on Philosophy and Computers*, and the future of the Barwise Prize. I am a big fan of academic freedom, and democracy in general, and so I think those issues should be put under broad consideration that goes far beyond membership of the committee. The role of the committee, and even more so of its chair, is to formulate productive topics for consideration, create the forum for stakeholder discussion, and listen to any conversations that may ensue.

1. SOME OF THE REASONS AND CAUSES.

1a. First, my take on the historical background.

Technological changes created groups of enthusiastic early adopters around the turn of the twentieth century in the areas such as computer programmers, web-development, online teachers, and so on. Early adopters were also the core of philosophers interested in computers, some of them—such as Jon Dorbolo, Robert Cavalier, Anthony Beavers, Ron Barnette, Bill Uzgalis, and Marvin Croy—have shaped, directly or through the CAP movement, this committee—at least for the first half a dozen years. There was much enthusiasm and people were devoted to *the cause*, though detailed understandings of the cause varied. The committee met regularly and the links between CAP and the committee were strong. The thriving of the committee can be seen in the spring 2004 issue of the newsletter, especially in the report from then chair Marvin Croy.[2]

The strong link between the committee and CAP resulted in the fact that the distinction between the two remained vague. There were worries at the APA that the committee was not truly open to the philosophers not associated with CAP. This resulted, around 2003-2004, in the influx of committee members unconnected with CAP—I think I actually benefited from this change by being able to join the committee on this wave. Those important and potentially beneficial changes gradually decreased cohesion of the group. Some of the members participated in no activities or were not available for physical meetings (much later this was the case even with some of the committee chairs). As can be seen on the committee's website, one of the chairs failed to even post obligatory reports throughout the whole term in office,[3] and one of these years we even failed to award the Barwise Prize (despite a valid vote taken by committee members). Also, the link with APA executive directors and persons responsible for the APA's electronic presence gradually weakened, although the committee always had some of the experts that could have been helpful in making decisions on the electronic technology and web presence. The original mission of the committee has been drafted as a rather basic document, so as to build a coalition even with those members of the APA board who, at the time, did not see computers as relevant for philosophy, except for typing their articles. There was always a feeling that we needed to wait a bit longer to show the committee's true colors.

I do not think we should belabor on the historical background, but some level of clarity is essential and further clarification, especially from the colleagues involved in the first years of the committee, would be very welcome.

1b. The mission of the committee.

Already during the chairmanship of Michael Byron, around 2008, we started getting encouragement from the APA leadership to update the committee's mission. This was not viewed as an urgent task, though it started some reflection. Much of it was lost during the following years. The second nudge from the APA leadership is visible in the committee

report for 2015-2016.[4] At the very end of his term as chair, Tom Powers received a clear message from the APA for the committee to update its mission, preferably including its name. It would have been an easy way to sneak out of the now endangered with extinction class of "philosophy of" committees and to align the written mission with what our real activities were. Unfortunately, proper attention to the mission statement was put second to day-to-day operation. Committee leadership (Marcello Guarini, then the chair, and I, then the vice-chair) revised the committee charges as late as 2018, which was after the APA Executive Committee resolved to discontinue it; as the saying goes, it was *too little too late*. As the 2017-2018 report indicates, those revised charges were "well received by the committee," which makes it a bit murky whether they have been formally adopted. Assuming the affirmative, committee charges now are as follows:

> The committee works to provide forums for discourse devoted to the critical and creative examination of the role of information, computation, computers, and other computationally enabled technologies (such as robots). The committee endeavors to use that discourse not only to enrich philosophical research and pedagogy, but to reach beyond philosophy to enrich other discourses, both academic and non-academic.

As one of the first steps as the chair, I submitted those charges to an up-or-down vote by the committee and they have been adopted unanimously. Those current charges are now on the agenda of the APA Board in its November meeting with our hope for approval. The reason for the up-or-down vote has been the lack of time for philosophical discussions on this. I do not view the above as perfect, but we have been trying to formulate the perfect mission statement for almost a dozen years and time has run out on us. The reason for working on the charges at all, the charges for a committee being closed down in a matter of months, is for the sake of clarifying what we represent, the issues we have developed or needed to be working on. For the most part, this definition is meant to be descriptive of what we have been doing at the committee sessions, in the newsletter, and in other ways. Some of those tasks need to be articulated in order to be explicitly taken over by the APA when the committee is gone, while others may need to be pursued by a follow-up group or groups after the committee's retirement. The above is just an introduction to the more practical discussion.

1c. The reasons for the Executive Committee's decision.

Back to the announcement from February 23, 2018. The first reason for closing the three "philosophy and" committees is that those committees were created to address pressing needs, and those needs no longer exist. The claim is addressed to the three rather different committees and so it is overly broad to allow for fruitful discussion. However, the second reason, which I list below, is addressed specifically to this committee.

The second reason is that the board understands the committee's mission to deal with "the use of computers by philosophers for instruction, writing, and publishing," which at the time the committee was created "was relatively unexplored territory." Is it a misunderstanding? Well, it may be a *de facto* misunderstanding of our activity, but—guess what?—this is what the committee's official charge was through 2017, and it was not corrected despite the committee being asked, at least twice, for a major update. For the sake of clarity, here is what the charge was in 2017:

> The committee (created by the board in 1985) collects and disseminates information on the use of computers in the profession, including their use in instruction, research, writing, and publication, and it makes recommendations for appropriate actions of the board or programs of the association.

The Executive Committee was right to take the charge of the committee at face value—the fact that it had not been revised properly lies on the committee, primarily on the committee's leadership for the last dozen years. Of course, every committee member, and especially myself as a long-standing member and newsletter editor, could have moved the mission changes forward—but we've failed. That's the answer to Fritz J. McDonald's well-meaning comments. The Executive Committee does not deal with Platonic images of the committee; it does not even evaluate it primarily based on the content of its sessions, newsletter, or oral testimonies. The Board of Officers is supposed to focus primarily on its reports and even more so on the mission statement. Organizations unable to pass a basic test of revising their antiquated mission statement are likely to be dysfunctional also in other ways, or so they seem.

2. SOME OF THE PRACTICAL ISSUES FOR THE REMAINING YEAR.

2a. Organizing sessions on philosophical questions in information technology at APA meetings.

Currently we have five proposals for the 2020 APA Sessions. The session "Philosophical Approaches to Data Justice," organized by Daniel Susser, has been accepted by the Eastern Division. The sessions "The Unreasonable Effectiveness of Logic in the Computational Sciences," organized by Gary Mar, and "Women in Tech: Things You Need to Know," organized by Susan Sterrett, have been submitted to the Central Division. A session titled "Machine Consciousness and Artificial General Intelligence" and a Barwise Prize award session are being finalized to be submitted for the Pacific Division.[5] With this level of interest in organizing solid sessions related to the committee's actual mission, it would be a waste to lose this capacity.

If there is a silver lining, it comes in here. In correspondence with my predecessor, Marcello Guarini, the APA offered to give the status of an affiliated group, if a group was built out of the current and former members or activists of this committee. As I understand, it is not trivial to gain such status. More importantly for the issue at hand, such a group has the right to propose a session for each of the APA divisional meetings. I understand that it should not apply for more than one session for each divisional meeting, and that such applications are prioritized just below those by

the committees. I think that this is a relatively good deal—perhaps too good to pass up.

However, we need a thorough, democratic discussion about whether to create an affiliated group. Even more importantly, we would need to define what such a group would need to focus on and who would want to give the time to develop it.

2b. Active legacy of the APA Newsletter on Philosophy and Computers.

It is clear that the *APA Newsletter on Philosophy and Computers* may not be published outside of the APA, not under this name. While *circa* 2012 we had a request from the APA to turn our newsletter into an APA journal, it was long before the APA established its official journal.

The issue at hand—and I am talking here with my hat as the newsletter's editor on—is to preserve and enhance the influence of our newsletter's legacy.

The *APA Newsletter on Philosophy and Computers* has played multiple roles.

A. Newsletter as a documentary of the committee's past.

It documented accomplishments of the committee, recorded its history, and recognized people active at the committee. This function was predominant during Jon Dorbolo's editorship (the first five years or so), but it has been a vital function of every issue of the newsletter—including the current issue. For this, it is important to gain the APA's commitment of keeping available the newsletter "forever," which in practical terms means, at least, while the organization exists. It would also be good to allow somebody, at least our "affiliated group" (should we create one), to mirror those newsletters on their website, as part of our shared legacy.

B. Newsletter as a repository of major philosophical masterpieces.

Several major philosophers have decided to publish their original articles with us, largely trusting that the name of the APA and the open access status of the newsletter would guarantee their work's survival and high visibility.

Those masterpieces include two original articles by Hintikka, organized by M. Kolak; an important paper by John Pollock, published posthumously, in our newsletter, by Terry Horgan, charged by Pollock's family to find the most appropriate place for this 53-page-long article (Terry also wrote a substantial introduction); an important article by Lynne Rudder Baker, with commentaries by Amie Thomasson and other top philosophers of the younger generation; original works by Gilbert Harman, Bernard Baars, Stan Franklin, Susan Stuart, Greg Chaitin, and many up-and-coming scholars; and, also, Barwise Prize winners such as Luciano Floridi (we published several of his articles since 2002, and a number of important commentaries on his work), J. Moor, T. Bynum, W. Rapaport, J. Copeland and G. Piccinini.

Those and many other outstanding articles need not only secure preservation but also promulgation. Many journals today, including the open access ones, help organize anthologies based on their content. It would be a great project to undertake, maybe by working directly with a publisher or to be undertaken by the new affiliated group together with the APA. Those topics should remain on the table.

Due to the changes in the manner in which APA Newsletters have been presented at the APA website, which took place *circa* 2013 (that eliminated webpages and kept the newsletters only as PDFs), currently the articles published in this and other newsletters are practically non-web-searchable. I have been working with my former office assistant on producing a list of all the articles published in the newsletter, which—if completed—may serve as the beginning of an easier-to-search catalogue.

C. Newsletter as a living journal.

Finally, there is a question of producing content of the sort this newsletter has been. A follow-up group may want to do so, without the APA affiliation. Within the APA, as Amy Ferrer recommended in a recent email, we "might consider working with the APA Blog to develop a periodic blog series—they do that for some committees and I expect would be willing to work with affiliated groups as well." Many other options exist as well, but starting a new journal by a different group may not rank high on the committee's busy agenda.

Again, we need a broad discussion among the many stakeholders—at the APA, in the committee, and out in the community—to work out the best ways to clarify and satisfy at least the top two of the above objectives. However, first, we want to work on building some approximation of a consensus on what "we" are going to be starting July 2020—and whether "we" want to be anything, as a group.

2c. Future of the Barwise Prize.

The Barwise Prize has been approved by the APA, at the request of CAP, as a unique committee-based APA prize. The APA does not mean to stop awarding it. It is meant to "officially be put under the oversight of the larger APA prize committee—the Committee on Lectures, Publications, and Research." (as stated by Amy Ferrer in a recent communication). Amy continues in the same message, "we will continue to ensure that appropriate specialist expertise is part of the selection process, and we can certainly discuss a role for the new affiliated group in that process. Perhaps the affiliated group could be given a set portion of the seats on the Barwise Prize selection committee, for example." This is a step in the right direction, and also an invitation for further discussion. This approach should be appreciated and acted upon by the committee.

Again, we should gather the relevant stakeholders. I think that primary group of stakeholders in a position to shape up the future of the Barwise Prize are the past winners of this prize. But, of course, the option of the "affiliated group" being given "a set portion of the seats on the Barwise Prize

selection committee" is very much worth keeping on the table.

2d. New proposals.

We should be very much open to new topics and initiatives. This should pertain to the current committee members, the ones whose term expires, and the broad community of stakeholders.

3. IMPORTANCE OF COMMITTEE ACTIVITIES AND MEMBERSHIP.

3a. Positive message from the APA National Office.

The quotes from Amy Ferrer that appear earlier in this note all come from her May 14, 2019, email in response to Marcello Guarini's question about the future of the newsletter and the Barwise Prize. It was addressed to Marcello in his capacity as the chair with me CC-d as the vice-chair. Thus, in my capacity as the current chair, I think it is my call to share its important parts with the committee and other stakeholders, even more so since the letter sends a positive message and opens up the space for further productive collaboration.

Hence, whatever the feelings of some of the stakeholders associated with the committee, I am willing to argue that the APA should still be viewed as a reliable partner, and to some degree potential home for several of the initiatives related to *philosophy and computers* that are currently carried on by this committee.

3b. Membership and the stakeholders.

Starting today, only Jack, Robin, Susan, Daniel, Gary, and I are official members of this committee. This is a small group of people. Hence, while we do not expect extraordinary feats from anybody—including the chair of this committee—we must expect of all the members to fulfill their duties. As I have learned last week, from an excellent online training for committee chairs, all members have an active duty to work for the committee, to a reasonable extent, and the committees have the option to ask the APA to replace those failing to do so.

Second, it is customary to ask active members of the cohort whose terms just expired to stay on the email list and continue with their ongoing tasks—in our predicament it is very important since this cohort is not being replaced. We should all volunteer what we are going to do for the committee and follow up on those promises.

4. SUMMARY—PREPARING FOR THE NOT-SO-DISTANT FUTURE.

First, I think, we want to build a common vision on what "we" are going to be after June 30, 2020—and whether "we" want to be anything, as a group. The committee needs to organize a dialogue with all the stakeholders this committee serves, such as past and present committee members, the Barwise Prize winners, participants in our sessions, authors publishing in the newsletter, readers, audiences, and many others. We need to see if they care to continue the workings of this committee, in a new venue, and what the follow-up activities would be. And if there is no interest, then, well, we would have done our due diligence and move on with our lives.

One final clarification, this is the note of the incoming chair, with my personal opinions and proposed projects. It has to be submitted on the first day of my term as the chair to the publication schedule at the APA. It will be consulted with the committee throughout the fall, and I am sure many improvements shall be made to those plans and ideas. But we need an action plan swiftly, and here is a draft. It is informed by my various roles on this committee for the last fifteen years, which may be an asset, but also a hindrance in designing truly new things. Hence the need for all committee members, as well as all the stakeholders for whom this committee operates, to address their visions, initiatives, and productive concerns.

Sincerely,

Piotr (Peter) Boltuc
Professor of Philosophy and Associated Faculty of Computer Science, University of Illinois Springfield; and University Professor of Online Learning, Warsaw School of Economics

NOTES

1. https://www.apaonline.org/news/388037/Changes-to-APA-Committees.htm
2. M. Croy, "From the Chair," *APA Newsletter on Philosophy and Computers* 3, no. 2 (2004): 2. Available at https://cdn.ymaws.com/www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/ComputersV03n2.pdf
3. https://www.apaonline.org/members/group.aspx?id=110436
4. https://www.apaonline.org/members/group.aspx?id=110436
5. Those are the titles and expectations as of July 2019. All of our proposals have been accepted. The list of the actual session titles, participants, and in most cases days and times of the sessions have been listed in the note from the editor and the announcements that appear at the end of this issue.

FEATURED ARTICLES

How Many Thoughts Can Fit in the Form of A Proposition?

S. G. Sterrett
WICHITA STATE UNIVERSITY

1. INTRODUCTION AND OVERVIEW

Let us agree on this much: people use sentences to communicate. On the view that sometimes it is thoughts that are communicated, then, sentences can be used to communicate thoughts. This was Frege's view. However, sentences are used not only to communicate thoughts, but to do other things as well. And, sometimes, in conversation and in writing in natural language, people rely on more than the sentence itself to communicate a thought. This, too, was Frege's view. The study of language is not the study of logic.

Still, developing a logic can start with the study of language, and progress by clarifying how logic is different from language. Unlike the study of language, logic studies sentences only inasmuch as they are used to communicate thoughts, and logic is about using nothing more than sentences to do so. This approach is sometimes part of an anti-psychologistic program in logic, for these two differences between language and logic involve separating the communication of thoughts from psychological aspects of communication.

It can be a bit misleading to call this approach anti-psychologistic without qualification, though, for it is not against the use of psychology in places other than logic: in fact, it draws attention to the fact that psychology is involved in many cases of human communication. An anti-psychologistic view of logic based upon a conviction that psychology has no place in logic was not the only motivation a mathematician might have for distinguishing logic from natural language. For, by Frege's time, it was becoming clear to many mathematicians that natural language, no matter how well-suited it might be for conversation, prose, and poetry, was not always up to the task of providing a language in which to prove theorems and show how the truth of one thought depends upon the truth of another. Relying on natural language, different mathematicians produced proofs whose conclusions were in conflict. What was needed was a means of referring to thoughts that allowed one to determine how thoughts were related to each other.

The anti-psychologistic aspect of Frege's approach is the conviction that relations between thoughts are not a matter of human psychology of any sort whatsoever, general or individual, and hence that any formal calculus of sentences meant to reflect the relationship between thoughts should not involve psychology. In Frege's writings, there was never any wavering, never the slightest hint of compromise, on this point.

Another, less crucial goal, was that the means of expressing a thought using a sentence in this formal calculus be at the same time a means of communicating thoughts that did not depend upon the contingencies of an individual human's psychology. This latter goal was something aimed at, but even Frege came to realize that there was no guarantee that it could be achieved in every case using the system of logical notation he had developed.

Sentences are used to enable people to grasp thoughts, but there can be cases where a sentence enables some people, but not others, to grasp the thought it expresses. Ultimately, communication of thoughts relies on some common understanding, including "a store of concepts" held in common; Frege wrote that a common store of concepts is handed down from one generation to the next,[1] and that learning a language necessarily involves nonverbal communication. Once the common understanding of relevant concepts is achieved, though, it becomes possible for communication of thoughts to proceed such that the sentence alone communicates the thought, without any reference to particulars such as the context in which it was uttered or particulars of the psychological state of the speaker or hearer. When, however, this common understanding is lacking, as Frege fretted might be the case with certain concepts such as the mathematical/logical notion of a "course of values" of a function used in his formulation of Basic Law V, Frege did not count on everyone grasping the thought expressed by the sentence. Hence different people might disagree on the truth of a sentence, yet their disagreement not be a matter of one person holding a grasped thought true and the other person holding the same thought false. The disagreement arises because they are not grasping the same thought.

That different people hearing or reading a sentence might not grasp the same thought by it was an unhappy situation that, as time went on, Frege learned to accept as inescapable in practice—in the general case. For the more specific and very urgent issue of giving a foundation for arithmetic, however, he at least began the project with the hope that the necessary common understanding might be achieved among mathematicians, if only the right formalism for expressing the thoughts used in arithmetic might be developed. That was the goal of the *Begriffsschrift* (Concept Script): to develop a formalism for statements that expressed the thoughts needed to prove theorems of arithmetic. It was not until he got down to working out the project in detail in the *Grundgesetze* (*Basic Laws of Arithmetic*) that he hit the little snags he regarded as temporary imperfections that might eventually be perfected.[2]

The formalism Frege developed in the *Begriffsschrift* is generally regarded as containing the fundamental features of modern symbolic logic—and so as an historically significant breakthrough from previous logics, including subject-predicate logic and Boolean logic. Thus reverential appellations such as "the father of modern logic" are heaped upon Frege. In light of such an accomplishment, many have wondered at Frege's seeming antipathy to formalism in arithmetic, and to his criticisms of Hilbert's formalization of geometry, especially his criticisms of implicit definitions of concepts such as the concept of point. Some sympathetic examinations of Frege's views have already been offered.[3] One point on which Frege did strenuously diverge from Hilbert was on the use of (what we would now call) uninterpreted statements as premises in proofs.

The thesis I will put forth in this paper—that, ultimately, Frege came to the view that "ideally, one sentence, one thought; one thought, one sentence"—bears on questions about how to explain and understand Frege's criticisms of formalism in arithmetic and Hilbert's use of uninterpreted statements and implicit definition. These criticisms Frege made of his contemporaries are not ad hoc reactions or piecemeal bits of philosophy, but reflect a more unified view about the relation between sentences and thoughts that was slowly being clarified as he tried to attain his original goals of putting mathematics on a solid foundation. The view about the relation between sentences and thoughts that developed over time was as much a result of frustrations encountered as of initial convictions vindicated, and one that he came to accept as practically necessary over time, rather than one chosen as a starting point.

2. LANGUAGE AND LOGIC

According to Frege, it is the thought that matters—at least to the logician. Sentences express thoughts, he said.[4] But he also found it frustrating to have to use sentences to communicate thoughts. He lamented that he could not put a thought in the hands of his readers as a mineralogist might put a rock-crystal in the hands of audience members. "One fights against language," he wrote in a footnote to his essay "Thought," and I am compelled to occupy myself with language although it is not my proper concern here."[5] According to him, the logician is concerned only with the thought expressed by the sentence. The thought, however, cannot be handled on its own; it can only be dealt with as wrapped in a linguistic form.

If sentences express thoughts, then what is the problem? Twofold: sometimes—often, in fact—(i) the content of a sentence goes beyond the thought it expresses. Sometimes "the opposite" happens, instead: (ii) the thought expressed goes beyond the content of a sentence; the "mere wording . . . does not suffice for the expression of the thought."[6]

The examples Frege gave of sentences in natural language in which the content of the sentence goes beyond the thought expressed by it (i.e., of case (i) above) consisted of pairs of sentences that expressed the same thought, although one sentence in the pair was a transformed version of the other. Although the sentences expressed the same thought, one of the sentences had a content that the other did not. An example is the following pair: "Alfred has not come" and "Alfred has not yet come"—the latter sentence differs from the former by the addition of the word "yet" and it creates expectations in the hearer that the first does not. Some other examples of transforming one sentence into another that expresses the same thought but has a different content are the following: (a) replacing "but" for "and" in a sentence, (b) adding "still" or "already" to emphasize part of a sentence, and (c) changing the verb from active to passive and the accusative into the subject.

In all these examples of case (i), the difference between the original and transformed sentences may not be trivial from the standpoint of what the hearer comes to understand or expect upon hearing it, so the sentences may well be said to differ. But, according to Frege's own remarks on such examples, the original and transformed sentence do express the same thought. This is because these transformations "do not touch the thought, they do not touch what is true or false."[7] Logical relationships are relationships between thoughts; the relationships between thoughts in which Frege was interested were relationships of one thought's dependence on another for justification. So, in the context of an endeavor in which we are concerned only with relationships between thoughts, rather than with expectations created in the hearer, we are not concerned with variations on a given sentence that do not affect its truth value. This point is made by saying that the logician is not concerned with the difference between two such sentences.

"Alfred has still not come," says Frege, is not false even if Alfred's arrival is not expected. "Alfred has still not come" is a different sentence than "Alfred has not come," but the two sentences express the same thought.[8] That is because a thought, for Frege, "is something for which the question of truth can arise at all."[9] That different sentences of our natural language can express the same thought is no problem for the logician: the logician just doesn't distinguish between them. In the *Begriffsschrift*, Frege had explained the purpose of his logical symbolism by comparing it to a specialized mechanical aid for seeing: a microscope. The comparison was meant to emphasize that logical symbolism is designed for a special purpose, as is a microscope. Logic is as poor a tool for capturing all the distinctions important to understanding conversation and poetry as is a microscope for viewing a landscape. The point here is that one must be modest about the aims of a particular logical symbolism. Or, rather, one must be clear about its purpose. For purposes of capturing what is relevant to the kind of content of a sentence of interest in doing logic, sentences that express the same thought should not be distinguished.[10] Let's call this activity—the activity of, for the purposes of logic, identifying sentences that express the same thought—pruning sentences.

What about the opposite situation, i.e., case (ii), when the thought expressed goes beyond the content of a sentence? To Frege, the task is clear: since logic is concerned only with thoughts, we need to augment such a sentence so that the thought determined by the sentence is unique. The kind of sentence that logic is concerned with is the kind that expresses a thought. Case (i) was the case in which different sentences may express the same thought—this the logician tolerates by not distinguishing between the various sentences, and perhaps selecting one as canonical and not using the others—but case (ii) is not so easily accommodated, for it is intolerable that the same sentence should express different thoughts.

Hence, when a sentence contains indexicals (e.g., "I," "this," "that," "yesterday") or proper names (e.g., Dr. Lauben, Venus) the logician is in trouble if stuck with only the sentence to go on. Thus, Frege says that "The words 'This tree is covered with green leaves' are not sufficient by themselves to constitute the expression of a thought, for the time of utterance is involved as well." He continues: "Only a sentence with the time specification filled out, a sentence complete in every respect, expresses a thought."[11] Let's call this activity extending a sentence, on analogy with a gardener's activity in doing the opposite of pruning a small offshoot—here, each of a set of multiple offshoots is encouraged to grow and extend itself to become a distinguishable branch in its own right.

What, then, he asks, should a logician make of a sentence such as "Dr Lauben was wounded"? This sentence expresses different thoughts to different people, depending upon the meaning each associates with "Dr Lauben," which in turn may depend upon whether they are acquainted with him and what they know about him. Frege gives a long example in which various people associate different definite descriptions with the proper name "Dr. Lauben," and are in different states of ignorance or knowledge about whether people with whom they are acquainted fit those descriptions. A fellow named Leo and a fellow named

Rudolph both hear Dr. Lauben say aloud, "I was wounded." Later, Rudolph hears Leo report aloud, "Dr. Lauben was wounded." Whether or not the statement made by Dr. Lauben and the statement made by Leo express the same thought to Rudolph is going to depend upon whether or not Rudolph knew that the man he heard saying, "I was wounded" was Dr. Lauben. The point of these examples is that "Dr. Lauben" is a proper name, but there may be different modes of determining the man to whom it refers ("the way that the object so designated is presented"). The logician does need to take these differences into account, Frege says, for different modes of determination for "Dr. Lauben" will result in different thoughts being expressed by the sentence "Dr. Lauben was wounded."[12]

Here, the problem is not a matter of difference in truth value of the different thoughts, for, says Frege, either the thoughts expressed by the sentence are all true or the thoughts expressed by it are all false. The problem is that knowledge of the truth of these thoughts can differ due to different hearers' mode of determination of the person to whom a proper name refers, and this indicates that the thoughts are different. In "Thought," Frege addresses this kind of case—i.e., the kind of case wherein the same sentence can be used to express different thoughts—by adding a restriction on sentences that will be permitted in a logical treatment of any topic involving proper names. The restriction is this: restrict the meaning (sense) of proper names so that no sentence expresses more than one thought. Let's call this kind of activity extending a sentence, too, for, as in the other examples of case (ii), it is analogous to dealing with multiple offshoots by encouraging each offshoot to take its own shape, and so distinguishing each offshoot from each other. However, we do not ever use more than one proper name for an individual—we may have multiple modes of determination that happen to determine the same object, but no proper name has as its meaning more than one mode of determination. In Frege's words: "So we must really stipulate that for every proper name there shall be just one associated manner of presentation of the object so designated. It is often unimportant that this stipulation should be fulfilled, but not always."[13]

3. NATURAL LANGUAGE AND THE FORMAL GARDEN OF PROPOSITIONS

Thus, Frege requires that the sentences of one's natural language that are the concern of logic be in some cases extended (distinguished from each other) and in some cases pruned (identified with each other) so that the relationships that hold between the resulting sentences—sentences the logician can, so to speak, hold in his hand and show to his audience—express the relationships that hold between the thoughts they express. As described in the previous section, sentences that express the same thought are not distinguished from each other (metaphorically, the several branches are pruned down to a single branch). Sentences that do not determine exactly one thought are extended (so that they determine only one thought) or disambiguated such that several sentences, each of which determines exactly one thought, are obtained.

The result is that, for the pile of sentences with which the logician deigns to work, each such sentence expresses exactly one thought, each thought is expressed by exactly one such sentence, and the relation of consequence between such sentences expresses the relation of consequence between the thoughts they express. Of course, this is not true for all the sentences of one's natural language—the point is that it is true of all the sentences the logician is working with after extending and pruning them per the prescriptions just described. Frege eventually came to see such prescriptions as necessary.

We can call the items that result from this process propositions, once they meet such prescriptions; it is irrelevant whether or not the resulting items also happen to be sentences of a natural language. In the *Begriffsschrift*, in explaining the value of the notation he introduced as a replacement for subject-predicate form, Frege said the symbolism he was presenting was a useful tool, if the task of philosophy was to "break the power of words over the human mind" and to free thought "from the taint of ordinary linguistic means of expression."[14]

Some readers may take issue with the point just made above, that for the pile of sentences with which the logician deigns to work, each thought is expressed by exactly one sentence, each sentence expresses exactly one thought, and the relation of consequence between sentences expresses the relation of consequence between the thoughts they express. I am well aware that not everyone who has encountered Frege's writings has the impression that Frege avoids the situation wherein a thought is expressed by more than one sentence. Nevertheless this is what Frege says in "Thought." He wrote "Thought" over twenty-five years after writing the much-emphasized and more widely studied "On Sense and Reference" and almost forty years[15] after the publication of *Begriffsschrift*, the work in which he introduced the formalism suitable for doing arithmetic in a "calculus of pure thought." In the *Begriffsschrift* (which predated a distinction he later drew between sense and reference), he did begin to lay out a view that was later revised. As I see it, the vision and ideal he had are not rejected, but rather are better realized, in the view he later laid out in "Thought." In "Thought" he explains more fully, and with examples, the process that I have referred to as the extending and pruning of sentences in the natural language required to obtain the kind of propositions that are fitting for the study of logic.

There's a similar progression in Frege's work concerning his attitude towards the relation between sentences and the thoughts they express. Frege's break with the traditional subject-predicate form of his predecessors, which he discusses in the *Begriffsschrift*, is accompanied by the statement in that early work that this break with tradition is warranted, "that logic hitherto has always followed ordinary language and grammar too closely."[16] In the much later "Thought," Frege writes that although he is not in the "happy position" of the mineralogist who can exhibit the gem he is talking about, he is resolved to a kind of resentful contentment: "Something in itself not perceptible by sense, the thought, is presented to the reader—and I must be content with that—wrapped up in a perceptible linguistic

form." It is not, however, a totally peaceful contentment: "The pictorial aspect of language presents difficulties. The sensible always breaks in and makes expressions pictorial and so improper."[17] The contentment he has achieved is the serenity of accepting what he cannot change.

Frege's explanation of this point—that there are differences between the linguistic forms one needs in natural language (where sentences have additional functions not relevant to logic, such as the function of generating expectations in a hearer that enable conversations to be carried on effectively, and the function of generating ideational associations), and the logical forms one needs to establish the truths of arithmetic—also illuminates his critique of Hilbert. For once one sees the view he expresses in "Thought" about the relationship between sentences and the thoughts they express as the view he was in the process of working towards when he responded to Hilbert's *Foundations of Geometry*, Frege's response to Hilbert's formalization of geometrical axioms seems quite natural.

Hilbert's axioms of geometry were (what we would call) uninterpreted: they were neither true nor false, until they received an interpretation. Frege's complaint was that the notion of an interpretation of a proposition was fundamentally incompatible with the notion of proposition required to do logic. It's easy to see why he thought so: logical relations hold between thoughts. A proposition—the kind of extended and pruned sentence logicians deal with—expresses a thought, and only one thought. On this view, the notion of interpretation has no place in logic.

In his correspondence with Hilbert, Frege wrote that "one feels the broad, imperspicuous and imprecise character of word language to be an obstacle, and to remedy this, one creates a sign language in which the investigation can be conducted in a more perspicuous way and with more precision." He used a slightly different horticultural metaphor, the process of lignification, to illustrate a point about symbolism: Instituting a new symbolism is like the tree's new growth hardening—after it has had a chance to take on the shape appropriate to performing its function.

Then, additional new growth depends upon those hardened sections to support the delivery of nutrients to the newly forming branchtips. The point is that trees do not grow into a predetermined suit of armor made of bark. The rigidity provided by the bark comes only after new branch tips have had a chance to grow in a natural formation. The sign language of a science is not set independently of inquiring as to what signs are best suited to it; if developed appropriately, these signs can be used to hook imperceptible thoughts and wrap them in a perceptible form so there is something that can be held in one's hand, so to speak, and worked with. Signs always involve a compromise compared to what one wishes to communicate, for, after all, signs are perceptible and the thoughts they express are not. It is fundamental to Frege's view that having the right formalism available is important to being able to capture the kinds of imperceptible thoughts in which one is interested. Frege's remark to Hilbert that the need for symbolism comes first, and only later the satisfaction of that need, reflects this conviction. In correspondence, Hilbert expressed agreement with this last statement.[18]

Hilbert used axioms as implicit definitions of the concepts contained in them, though, and Frege didn't like that any more than he liked the fact that Hilbert's axioms required interpretation in order to express a thought. However, as critical as Frege might have seemed of Hilbert, he did evaluate Hilbert's formalization of geometry with the idea of showing how one might achieve what Hilbert was after in a proper manner.

In fact, he outlined a way to make sense of Hilbert's method of showing axioms independent of each other. Frege's reconstruction of Hilbert's independence proofs, however, only work for (what Frege called) real propositions, which Hilbert's axioms were not.[19]

Frege's method works as follows: one maps ("set(s) up a correspondence between") words of a language (in which, of course, the reference of every word is fully determinate) onto other words of the same language, subject to some restrictions. These restrictions include mapping proper names to proper names, concept-words to concept-words of the same level, and so on. The signs whose references belong to logic (e.g., negation, identity, subsumption, and subordination of concepts) are not mapped to different signs. Then, one can show that a thought G is independent of a group of thoughts β, if one can obtain from β and G, respectively, a map to a group of true thoughts β′ and a false thought G′. In Sterrett 1994 I argued that this was in fact somewhat like the approach Hilbert actually took, and so it was striking that Frege distinguishes his method from methods that employ interpretations of statements. The significance of the difference between Hilbert and Frege, I concluded there, had to do with differences in their accounts of how words come to mean what they do. I will not repeat that discussion here, as it is readily available elsewhere.[20]

It should be clear by now that Frege is not drawing a distinction between referring to a thought and referring to the perceptible linguistic form in which the thought is wrapped. Frege's point was that the only way he's got to show anyone what thought he is referring to is by wrapping it in a perceptible linguistic form. Hence in talking of the thought G′ to which G is mapped (via the mapping of words of a language as outlined above), Frege can hardly be talking about making substitutions of words in, and obtaining transformations of, anything other than sentences. Not just any old sentences of a natural language, however. These sentences or propositions are the result of extending and pruning sentences of the natural language so that each proposition expresses one and only one thought, and so that propositions that express the same thought are not distinguished from each other. I use the term "proposition" here because Frege isn't including all sentences of natural language. He doesn't talk about the forms he has to wrap thoughts in other than as the forms in which the thoughts are wrapped; these forms are not self-subsistent. He was certainly against the idea of developing symbolic forms first and then looking for thoughts that might fit into them. And I don't think he ever meant to talk about these symbolic forms other than as used to express thoughts.

Thus, for Frege, the notion of logical consequence arises for relationships between the imperceptible thoughts that are wrapped up in perceptible forms, not to the forms of the wrappers themselves. One cannot communicate thoughts except by capturing them in such a perceptible wrapping, so proofs and derivations proceed by way of rules that apply to propositions or statements. However, these propositions always express a thought: they are never empty wrappers. They are not in need of interpretation.

4. DEPARTED THOUGHTS

Hence, Frege says that if by sentence is meant the "external, audible, or visible that is supposed to express a thought," then it does not make any sense to say that one sentence is independent of another. The context in which Frege wrote this was in arguing that Hilbert had erred in the specific way he had gone about trying to establish the independence of the parallel postulate from the other axioms of geometry. It was in this context that Frege said that Hilbert makes a mistake in calling anything "the axiom of parallels," for, as Frege put it in the passage quoted above, it is not the same in every geometry: "Only the wording is the same; the thought-content is different in each particular geometry."[21]

Frege means here to warn against mistaking the "external, audible, or visible that is supposed to express a thought" for the thought. Logic is concerned with thoughts and how they are related to each other. So there is a realm of thought: it cannot be perceived by the senses, but it is like perceptible things in that it does not need an owner, as ideas do.[22] Frege's favorite example of a thought in his essay entitled "Thought" is the Pythagorean theorem. Different people can grasp the thought, and it can be communicated by wrapping it in a perceptible linguistic form.

But I don't think Frege intends to alert the reader to the existence of a logical calculus of "the external, audible, or visible that is supposed to express a thought." This would be a study of the relationships of linguistic forms, something Frege thought of interest for many purposes—understanding conversations and writing poetry, for instance—but decidedly not the subject matter of logic. Logic is about thoughts, it is about the laws of thought, the laws of the laws of science. It is about deriving proofs so that we can see how one thought depends upon another. It involves the linguistic forms in which these thoughts must be wrapped in order to be communicated, but only in the context of investigating which thoughts depend upon which other thoughts. It is not about relationships of dependence between perceptible linguistic forms. If there are such things as forms that exist as shed snakeskins left behind from departed thoughts, they are not the concern of logic; they are not the items of a calculus of pure thought.

For Frege, there is no such thing as a realm of linguistic forms within which no thoughts are wrapped but which are related to each other in virtue of their form by logical laws. The logical relations are not logical relations between linguistic forms.

5. THE UNITY OF THOUGHT AND EXPRESSION

Frege did discuss examples of different sentences that expressed the same thought, even in "Thought," arguing that "the content of a sentence often goes beyond the thought expressed in it." But his response to this observation was not to posit a new kind of logical law or a new kind of logical relation to account for how such sentences were related. In "Thought," he did not regard such situations as puzzles; he did not then consider them relevant to logic. Rather, his response to this observation about natural language was that the logician does not distinguish between such sentences.

Was Frege this blasé about different sentences that express the same thought because, on a view sometimes attributed to Frege, he thought that there are really two distinct things, sentences and thoughts, and thus that the distinction between sentences is a distinction that can be made only in the realm of what is derivable, and not in the realm of what is provable? I don't think that this is how Frege's views on sentences that express the same thought ought to be viewed.

Recall that what Frege said about pairs of distinct sentences that express the same thought was only that some such transformations between sentences must be recognized as admissible. But this wasn't a matter of recognizing relationships that obtain in a realm of equipollent propositions. In his 1906 letter to Husserl, in fact, Frege suggested that equipollent propositions could all be communicated by a single standard proposition.[23] In closing the letter, he remarks that the question of whether equipollent propositions are congruent "could well be debated for a hundred years or more." But he isn't concerned about the answer; he writes, "I do not see what criterion would allow us to decide this question objectively. . . . But I do find that if there is no objective criterion for answering a question, then the question has no place at all in science."[24] Placing significance on the difference in the relations that hold between sentences and the relations that hold between thoughts is attributing significance to exactly what, I think, he actually said ought to be de-emphasized.

We have seen that what Frege said about sentence transformations that do not affect the thought expressed was that sentences with differences that don't affect the thought expressed don't need to be distinguished when doing logic. All that the existence of transformations that yield two or more sentences expressing the same thought means to the logician is that, if propositions or statements admit of such transformations, one must recognize as admissible those transformations that do not affect the thought expressed. Once we see this point of Frege's, the apparition of the notion of derivability according to which things are not always as they seem disappears: i.e., the notion of derivability on which a thought when wrapped in a different wrapper might have different derivability relations disappears. There is a realm of thoughts (thoughts are not the property of individuals as ideas are, but they are not perceptible either), and it is distinct from the realm of perceptible things.[25] Logical laws are used in showing the relationships that exist between thoughts, via a proof. Thus, the realm in which logical rules apply involves both of these realms, since it includes both thoughts and signs; and this in turn is due to the unavoidable situation that

communication of thoughts requires that thoughts be wrapped in perceptible forms.

What about Frege's statement that a thought can be "carved up" in different ways? Doesn't the fact that the same thought could be carved up in different ways mean that the same thought could be expressed by different sentences? Yes and no—the difference being a matter of which language you are talking about. In natural language: Yes, the same thought can be expressed by different sentences that analyze the thought into subject and predicate differently; typically this will happen whenever the same sentence is transformed from the active to the passive voice. But in the formal language of the *Begriffsschrift*, the answer is no: the carving really captures the structure of the thought relevant to the kind of inferences one wants to be able to draw. That is, the whole point of the *Begriffsschrift* was that subject-predicate logic did not get at the structure of thought relevant to making inferences! In contrast, the formalism of the *Begriffsschrift* was created to ensure that all of the structure relevant to making inferences that were a matter of pure logic could be expressed.

The point that the situation of having only subject-predicate logic available is restrictive in spite of allowing many options might be explained using the metaphor of a plant, as follows: that situation is like having only a certain kind of analysis of the plant available to you, for instance, having only the option of describing a plant in terms of dividing it up into the edible food it bears and the part of the plant that produces the edible food. What's limiting about this is not a matter of how many ways there are to carve up the plant, for in fact the edible-food and plant-that-produces-food way of carving up a plant permits many different ways of carving up the plant. Depending upon what part of the plant a creature is interested in consuming, one could analyze the plant into an edible product and the remainder of the plant that produces it in different ways, just as the subject-predicate form allows one to express a thought in different ways depending upon what one chooses as the subject of the sentence. Rather (using the plant metaphor) the limitation is this: the available ways of analyzing the plant does not necessarily allow us to analyze the plant structure in the way required for investigations in natural science.

Analysis of a plant based on edible parts of the plant does exhibit something about the structure of the plant, of course, but it also obscures some of the structure of the plant. What we want is a general method of carving up the plant in a way that allows the flexibility and precision to exhibit various kinds of structure in the plant, a way that permits the many different kinds of carving ups of the plant needed for making inferences we want to draw to conduct research about a variety of questions that interest us.

On this analogy, what's wrong with subject-predicate logic is that the kinds of "carving up" of a thought it permits—and it may permit a number of alternatives—might not include the structure of the thought that is relevant to making the kinds of inferences in which one is interested. In contrast, the formalism of the *Begriffsschrift*, in which concepts are modeled on functions, is meant to introduce a formal language in which one can carve a thought in any way needed for making scientific inferences. The formalism provided in that work is supposed to be enough to permit making any inferences that are a matter of pure logic. This is not to say that the kind of structure sought for even when using the formalism of the *Begriffsschrift* may not be relative to the kinds of inferences one is interested in making (hence the formalism needed for chemistry and physics is left open in the *Begriffsschrift*; in my biological metaphor, the added formalism needed to carve the plant into its relevant parts might be the gene concept). It is to say that the formal language does not, as subject-predicate logic does, limit one to carving the plant into two parts according to a criterion that may never permit one to delineate the structure of the plant relevant to the inference in which one is interested.

The advance Frege offered was not a way of dissecting a thought into formalism and unformed thought, but, rather, consisted in a formalism that permitted carving thoughts in more useful ways than previous formalisms allowed. The separation of thought from sentence underlying the distinction between provability and derivability is not something we find in Frege. To describe such a disconnect as part of Frege's view misdescribes Frege's notion of a proposition in the same way that Aristotle's notion of form would be misdescribed by using Plato's notion of form. That is, in Plato's philosophy, forms exist in a realm separate from the things of which they are forms. Aristotle, too, used a metaphor from biology to break from Plato: that there are male and female animals, he said, does not imply that male and female exists as something separable from male and female animals.

To use another metaphor: in a certain science fiction television series, there is a creature that can transform itself into various shapes, called a shape-shifter. These shape-shifters can separate from their shapes and meld together somehow in a realm in which they are shapeless. But this is, after all, fiction. To make the metaphor of shape-shifters who take on various shapes fit Frege's account of sentences as the forms within which thoughts are wrapped, let us leave the details of this particular science fiction story behind and stipulate that shape-shifters take on human forms, that a given shape-shifter cannot take on every form, and, in fact, that the forms a particular shape-shifter takes on are not taken on by any other creature. (This corresponds to Frege's requirements that, in his formalism, thoughts are expressed by sentences, that more than one thought is expressible, and that no sentence expresses more than one thought.) Clearly, once we've figured out the shapes between which a particular shape-shifter can transform itself, we no longer need distinguish between those shapes.

The analogy to thoughts and the linguistic forms they take on is this: just as, in the science fiction story, a creature is apprehended via the senses by its sensible form, so a thought is expressed via a sentence. That, in effect, is Frege's unperturbed response in his essay "Thought" to the examples in which there are several sentences that express the same thought, such as two sentences that differ only in the manner used to designate an object. That is, in contrast

to the view that Frege is saying that there are two different calculi, one for thoughts and one for the linguistic forms in which they can be wrapped, Frege shows that he intends to avoid such commitments by stipulating that, when proper names are used, only one manner of presentation (e.g., for Venus, either "the morning star" or "the evening star," but not both) be permitted. Thus I do not think that, as is often supposed, Frege developed a calculus of sentences associated with something called derivability in addition to the calculus of thoughts associated with provability. His remark to Husserl (quoted earlier) that he does not think there is room in science for the question of whether equipollent propositions are congruent bears this out.

Looking back from the present, some people attribute to Frege's *Begriffsschrifft* the achievement of having developed a calculus of sentences related by derivability, which are accurately described in modern parlance as syntactic relations. This is not so, and Frege is explicit enough about what he was doing to make that clear. Frege's *Begriffsschrift* was to be a calculus of thoughts. There were reasons that the calculus had to involve symbolic formalism—to clarify thoughts, and to express them—but the calculus was not a "topic-neutral" calculus of symbolic or syntactic forms. That may be what a modern logician sees in looking at the *Begriffsschrift*, but it doesn't sound much like Frege's description of the *Begriffsschrift*. What it does sound like, however, is Frege's description of Leibniz's vision, which, he said, "was too grandiose for the attempt to realize it to go further than the bare preliminaries." (in Beany, p. 50)[26] Frege thought Leibniz's vision of a universal calculus an excellent guiding vision, but what he said about his own achievement in the *Begriffsschrift* with respect to Leibniz's visionary aim was that "even if this great aim cannot be achieved at the first attempt, one need not despair of a slow, step by step approach." The project, Frege said, "has to be limited provisionally" at first. And he identified the *Begriffsschrift* as one of the "realizations of the Leibnizian conception in particular fields."[27] He spoke of additions that would have to be made to extend it to geometry and then to the pure theory of motion, then mechanics, and then physics. These latter fields involve natural necessity as well as conceptual necessity.

In his correspondence with Hilbert, Frege writes that he thinks Hilbert is (mistakenly) treating geometry as if it were like arithmetic. Frege thought it an error to regard geometrical knowledge as having the same kind of basis as arithmetical knowledge. This is important, for it meant that Frege didn't think sentences or propositions of geometry were related to each other in the same way that statements of arithmetic were. The *Begriffsschrift* was to help in showing that arithmetical truths were truths of logic, but even this does not mean that the rules in the *Begriffsschrift* applied to topic-neutral sentences, for Frege did not take a formalist approach to arithmetic either. What I mean by this is that he did not allow (what we would now call) uninterpreted statements of arithmetic any more than he did statements of geometry. In the *Basic Laws of Arithmetic*, he reiterates his requirement on axioms, i.e., that all the terms in them must be defined. That he is not always able to meet the requirement should not be cited as evidence that some of the concepts are implicitly defined or are uninterpreted and to be interpreted at a later date. Rather, Frege's explanation of such undefined concepts is found in a statement he makes in preliminary remarks in the *Basic Laws*: "It will not always be possible to give a regular definition of everything, precisely because our endeavor must be to trace our way back to what is logically simple, which as such is not properly definable. I must then be satisfied with indicating what I intend by means of hints."[28] The principles of the *Begriffsschrift* may apply to every science, but according to Frege they do not include all the principles nor, even, all the formalism needed to do geometry, kinematics, physics, or chemistry. These await future development, he said.

Thus, we must avoid the anachronism of splitting asunder a propositional form from a thought. For Frege, a proposition is a thought wrapped in a perceptible linguistic form, i.e., a propositional form. The perceptible linguistic form it is possible to wrap a thought in may not be uniquely determined for a given thought, but the thought must be wrapped in some perceptible linguistic form or other. Hence the proposition cannot survive such a dissection. Even in developing a calculus in which the ideal is "one proposition, one thought; one thought, one proposition," a thought and its expression are not split apart. Throughout his correspondence with Hilbert, Frege seems concerned to speak of the proposition as a whole, i.e., a "real" proposition expressing a thought. The kind of axioms Hilbert proposed, which were neither true nor false, and so which, on Frege's view, did not express thoughts, were not, on his view, proper subjects of logic.

On Frege's view, a thought is necessarily wrapped in linguistic form if it is to be communicated, studied, or used in reasoning. Thoughts are individuals for Frege, somewhat as trees and humans were individuals for Aristotle. Aristotle was concerned (at least in some of his works) to hold out for the identity of an individual in spite of the different things that could be predicated of it, but in a way that didn't call for dissecting that individual into a self-subsistent form and something else. Similarly, what Frege thought was called for with respect to thoughts was a method of expressing an individual thought that exhibited the structure of the thought in such a way that we could see its relation to other thoughts, but in a way that didn't call for dissecting it into a self-subsistent linguistic part and something else. Frege also seemed to recognize different kinds of relations between thoughts, that the relations that were crucial might be different for different investigations and different disciplines. The *Begriffsschrift* was meant to provide a calculus in which to express thoughts that met the needs of the discipline of logic, i.e., a calculus in which the logical relations between thoughts would be exhibited. Frege continually warned against the tendencies of some of his contemporaries to take the approach of attempting to separate the propositional form of a proposition from the thought it expresses and treat it as self-subsistent. The admonition to refrain from attempting such fatal dissections, though, is quite general. It is as old as Aristotle and as new as post-analytic philosophy.

NOTES

1. G. Frege, *Philosophical and Mathematical Correspondence*, 59.
2. Here I am referring to Frege's remark in the *Grundgesetze*: "A dispute can arise, so far as I can see, only with regard to my Basic Law concerning courses-of-values (V), which logicians perhaps have not yet expressly enunciated, and yet is what people have in mind, for example, where they speak of the extensions of concepts." Frege, *The Basic Laws of Arithmetic. Exposition of the System*, 3-4. Here he does express confidence that the concept of course of values might be enunciated more clearly, and that, when it is, disputes about Basic Law V will be settled. His attitude towards the very different kind of problem later pointed out by Betrand Russell was not one of confidence in overcoming it, and I am not referring to Russell's paradox when speaking of "little snags." Frege addressed Russell's paradox in Appendix II to volume II of the *Grundgesetze*.
3. M. Resnik, "The Frege-Hilbert Controversy"; S. G. Sterrett, "Frege and Hilbert on the Foundations of Geometry"; P. Blanchette, "Frege and Hilbert on Consistency"; A. Antonelli and R. May, "Frege's New Science"; J. Tappenden, "Frege on Axioms, Indirect Proof, and Independence Arguments in Geometry: Did Frege Reject Independence Arguments?"
4. Frege, "Thought," 328.
5. Ibid., 329-30.
6. Ibid., 331.
7. Ibid.
8. Ibid.
9. M. Beany, *The Frege Reader*, 328.
10. In a 1906 letter to Husserl, Frege wrote that, while it is not possible to say exactly when two propositions are merely equipollent and when they are congruent, this is not an obstacle in principle: "All that would be needed would be a single standard proposition for each system of equipollent propositions, and any thought could be communicated by such a standard proposition. For given a standard proposition everyone would have the whole system of equipollent propositions, and he could make the transition to any one of them whose illumination was particularly to his taste." Beany, *The Frege Reader*, 303.
11. Frege, "Thought," 343.
12. Ibid., 333.
13. Ibid.
14. Beany, *The Frege Reader*, 51.
15. The *Begriffsschrift* (Concept-Script) was published in 1879; *Grundlagen der Arithmetik* (Foundations of Arithmetic) in 1884, "On Sense and Reference" in 1892, and "Thought" in 1918. "On the Foundations of Geometry" and associated correspondence with Hilbert and others was written around 1900.
16. Beany, *The Frege Reader*, 51.
17. Frege, "Thought," 334.
18. Hilbert 4.10.1895 in Frege, *Philosophical and Mathematical Correspondence*, 34.
19. Sterrett, "Frege and Hilbert on the Foundations of Geometry."
20. Giving a brief description of the contrast between Frege's account of elucidation and Hilbert's account of implicit definition risks mischaracterizing Hilbert as more formalist than he was, so I refer the reader to my discussion in Sterrett, "Frege and Hilbert on the Foundations of Geometry" in which I distinguish the positions of Hilbert, Korselt (who responded to Frege on Hilbert's behalf), and Frege. The paper is available free online at the Philosophy of Science Archives server, at http://philsci-archive.pitt.edu/723/.
21. Frege's analysis is that the fault lies in confounding first- and second-level concepts, such as the concept of point. There may be different first-level concepts of point, under which points fall: the Euclidean point-concept is one such first-level concept. If one likes, one may also define a second-level concept, within which the Euclidean point-concept and other first-level concepts, fall. A fuller discussion of Frege's point is given in Sterrett, "Frege and Hilbert on the Foundations of Geometry," 9.
22. Frege, "Thought," 337.
23. Beany, *The Frege Reader*, 302.
24. Ibid., 305.
25. Frege, "Thought," 337.
26. Beany, *The Frege Reader*, 50.
27. Ibid.
28. Frege, *The Basic Laws of Arithmetic. Exposition of the System*, 32.

REFERENCES

Antonelli, Aldo, and Robert May. "Frege's New Science." *Notre Dame Journal of Formal Logic* 41, no. 3 (2000): 242–70..

Beany, Michael, ed. *The Frege Reader*. Oxford: Blackwell Publishers, 1997.

Blanchette, Patricia. "Frege and Hilbert on Consistency." *Journal of Philosophy* 93 (1996): 317–36.

Frege, Gottlob. "Frege to Hilbert 6.1.1900." 1900. In Frege (1980), 43–48.

———. "Thought." 1918. In Beany (1997), 325–45.

———. *The Basic Laws of Arithmetic. Exposition of the System*. Translated and edited, with an introduction, by Montgomery Furth. Berkeley: University of California Press, 1964.

———. *Philosophical and Mathematical Correspondence*. Edited by Brian McGuinness and translated by Hans Kaal. Chicago: University of Chicago Press, 1980.

Resnik, Michael. "The Frege-Hilbert Controversy." *Philosophy and Phenomenological Research* 34 (1994): 386–403.

Ricketts, Thomas. "Frege's 1906 Foray into Metalogic." *Philosophical Topics* 25 (1997): 169–88.

Sterrett, S. G. "Frege and Hilbert on the Foundations of Geometry." 1994 Talk. Available on the Philosophy of Science e-print archives at http://philsci-archive.pitt.edu/723.

Tappenden, Jamie. "Frege on Axioms, Indirect Proof, and Independence Arguments in Geometry: Did Frege Reject Independence Arguments?" *Notre Dame Journal of Formal Logic* 41 (2001): 271–315.

Consciousness, Engineering, and Anthropomorphism

Ricardo Sanz
UNIVERSIDAD POLITÉCNICA DE MADRID

1. INTRODUCTION

The construction of conscious machines seems to be central to the old, core dream of the artificial intelligence community. It may well be a maximal challenge motivated by the pure *hybris* of builders playing God's role in creating new beings. It may also just be a challenging target to fuel researchers' motivation. However, we may be deeply puzzled concerning the reasons for engineers to pursue such an objective. Why do engineers want conscious machines? I am not saying that engineers are free from hubris or not in need for motivation, but I question if there is an *engineering reason* to do so.

In this article I will try to analyze such motives to discover these reasons and, in this process, reveal the excessive anthropomorphism that permeates this endeavor. Anthropomorphism is an easy trap, especially for philosophers. We can see it pervasively tinting the philosophy of consciousness. However, in the modest opinion of this engineer, philosophy shall transcend

humanism and focus on universal issues of value both for animals and machines.

2. THE ENGINEERING STANCE

The construction of intelligent machines is the central activity of control systems engineering. In fact, the core focus of activity of the control systems engineer is the design and implementation of *minds for machines*. For most people involved in cognitive science, saying that a PID controller is a mind is not just an overstatement; it is, simply, false. False because such a system lacks emotion, education, growth, learning, genealogy, personality . . . whatever.

This analysis of what minds are suffers from biological chauvinism. Anthropomorphism is pervasive in cognitive science, artificial intelligence, and robotics. This is understandable for historical reasons but shall be factored-out in the search or core mechanisms of mind.

A central principle of engineering is that systems shall include what is needed and only what is needed. Over-engineered systems are too complex, late delivered, and uneconomical. This principle shall also be applied to the endeavor of building conscious machines.

2.1 THE SYSTEMS ENGINEERING VEE

The systems engineering lifecycle (see Figure 1) starts with the specification of needs: what does the user of the system need from it. This need is stated in the form of a collection of user and system requirements.[1] The verification of satisfaction of these needs—the system validation for acceptance testing—is the final stage of the engineering life-cycle.

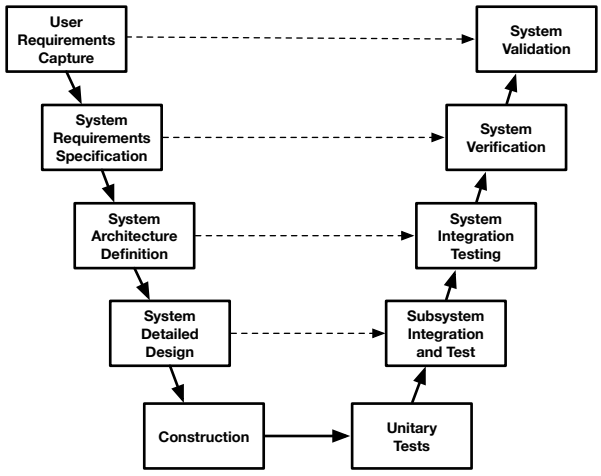


Figure 1. The Systems Engineering (SE) Vee. This flowgraph describes the stages of system development as correlated activities oriented to the satisfaction of user needs.

This process implies that all system elements—what is built at the construction stage—do always address a user or system need; they always have a function to perform. The concrete functions that are needed will depend on the type of system and we shall be aware of the simple fact that *not all systems are robots*. Or, to be more specific, not all intelligent systems are humanoid robots. Many times, intelligent *minds* are built for other kinds of systems. Intelligence is deployed in the sophisticated controllers that are needed to endow machines with the capability to address complex tasks. Minds are just control systems.[2] Intelligent minds are sophisticated control systems.[3]

2.2 NOT ALL AI SYSTEMS ARE ROBOTS

The obvious fact that not all AI systems are humanoid robots has important implications. The first one is that not all systems perform activities usually done by humans and hence:

1. Their realizations—their bodies—do not necessarily resemble human bodies. In engineering, bodies follow functional needs in a very intentional and teleological sense . Machines are artificial in the precise sense clarified by Simon.[4]

2. Their environments—the context where they perform the activity—are not human environments and fitness imply non-humanly capabilities.

3. Their missions—what are they built for–are sometimes human missions, but mostly not. People are worried about robots getting our jobs but most robot jobs cannot be performed by humans.

In control systems engineering we usually make the distinction between the controller—the mind—and the plant—the body. This may sound kind of cartesian and indeed it is. But it is not due to a metaphysical stance of control engineers but to the more earthly, common practice of addressing system construction by the integration of separately built parts.[5]

The plant (usually an artefact) can hence be quite close or quite different from humans or from animals:

Airplanes: Share the environment and the activity with birds, but their functional ways are so different from animals that control strategies are totally different.

Industrial Robots: In many cases can do activities that humans could do: welding, picking, packaging, etc. but requirements may be far from human: precision, speed, repeatability, weight, etc.

Vehicles: Autonomous vehicles share activity with animals: movement. However they are steadily departing from animal contexts and capabilities. Consider, for example, the use of GPS for autonomous driving or vehicle-to-infrastructure communication for augmented efficiency.

Chemical Plants: Some artefacts are extremely different from humans seen as autonomous entities moving in environments. Industrial continuous processes—chemical, oil, food—do not resemble humans nor animals and the needs for intelligence and awareness are hence quite different.

Utilities: The same can be said for technical infrastructure. The intelligence of the smart grid is not close to animal intelligence.

All these systems "live" in dynamic contexts and their controllers shall react appropriately to changing environmental conditions. They process sensory signals to be "aware" of relevant changes, but they do it in very different ways. Machines are not animals nor in their realization nor in their teleology. Bioinspiration can help systems engineers in the provision of architecting ideas of concrete designs of subsystems. However, mapping the whole iguana to a machine is not a sound engineering strategy.[6]

2.3 THE AI PROGRAM VS. THE STANDARD STRATEGY

From my perspective the many threads of the global AI program can be categorized into three basic kinds of motivations:

- Technology. Solving problems by means of incorporating intelligence into the artefacts.
- Science. Explore the nature of (human) intelligence by creating computer models of psychological theories.
- Hubris. Create beings like us.

Control system engineering (CSE) implements AIs because it is interested in the problem-solving capabilities that AI can provide to their machines. AI enters the CSE domain to deal with runtime *problems of higher complexity* that are not easily addressable by more conventional means. The mind of the machine is built as a cognitive agent that perceives and acts on the body that is situated in an environment. In their well-known textbook Russell and Norvig even say that "the concept of a controller in control theory is identical to that of an agent in AI."[7]

AI-based controllers can decide in real-time about what to do in complex situations to achieve system goals. Goals that are established in terms of user needs and, secondarily, in terms of machine needs. An AI-based controller does not pursue the machine objectives but the objectives of its owner.[8]

The optimal strategy for controlling a system is to invert a perfect model of it.[9] But this only works to the extent that the model behavior matches system behavior. Model fidelity is limited due to several factors. Observability limits what the intelligent agent may perceive. Note that the intelligent agent interacts with a body that interacts with the environment. In this situation perfect knowledge is unachievable because uncertainty permeating both the plant and its environment affects the intelligent agent mental representation (cf. the problems surrounding the deployment of autonomous cars).

Agent mental complexity shall match that of the plant and its environment following Ashby law of required variety. This implies that the *curse of complexity and uncertainty* affects not only the plant and its environment, but *also the controller itself*. Intelligent, autonomous controllers are enormously complex artefacts.

Complexity plays against system dependability. The probability of failure multiplies with system complexity. This is not good for real-life systems like cars, factories, or gas networks. The basic method to improve dependability is building better systems. Systems of better quality or systems built using better engineering processes (e.g., as is the case of cleanroom engineering). However, these do-well strategies are not easily translatable to the construction of systems of required high complexity.

The *standard strategy* to address runtime problems is to use humans to directly drive or supervise the system. Humans are better at addressing the unexpected and provide augmented robustness and resilience (R&R). A term that is gaining acceptance these days is Socio-Cyber-Physical System, *a system composed of physical bodies, software controllers and humans*. Figure 2 shows a common layering of these systems.

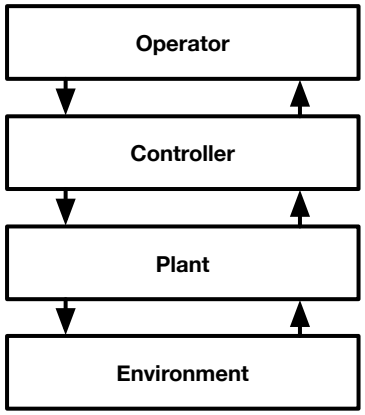


Figure 2. A socio-cyber-physical system is a layered structure of interacting systems. The top authority corresponds to human operators because they are able to deal with higher levels of uncertainty.

In socio-cyber-physical systems the top authority corresponds to human operators because they are able to deal with higher levels of uncertainty. Humans are able to understand better what is going on, especially when unexpected things happen. The world of the unexpected has never been a friendly world for AIs.

3. BUILDING CONSCIOUS MACHINES

The research for consciousness in artificial systems engineering can be aligned with the three motivations described in the previous section—useful technology, psychological science, or mere hubris.

Some authors consider that biological consciousness is just an epiphenomenon. However, an evolutionary psychology dogma states that *any currently active mental trait that has been exposed to evolutionary pressure has adaptive value*. This—in principle—implies that *consciousness has adaptive (behavioral) value*; so it may be useful in machines.

From a technological stance, the analysis/evolution of complex control systems took us into researching novel

strategies to improve system resilience by means of self-awareness. If the system is able to perceive and reason about its own disturbances, it will be able to act upon them and recover mission-oriented function. Machine consciousness enters the engineering agenda as a possible strategy to cope with complexity and uncertainty.

A conscious machine can reflect upon itself and this may be a potential solution to the curse of complexity problem. So, the engineering interest in consciousness is specifically focused on one concrete aspect: self-awareness. This implies that the engineering stance does not have much to say about other aspects of consciousness (esp. qualia).

3.1 SELF-AWARENESS IN MACHINES

Self-aware machines are aware of themselves. Self-awareness is just a particular case of awareness when the object of awareness is the machine itself. Self- awareness is a class of perceptual process, mapping the state of a system—the machine itself—into an exercisable representation.

From an engineering perspective, self-awareness is useless unless it is accompanied by concurrent action processes. In particular, to be of any use concerning system resilience, self-awareness processes need coupled self-action processes. This closes a control loop of the system upon itself.

This may sound enormously challenging and innovative, but this is not new at all. Systems that observe and act upon themselves have been common trade for decades. There are plenty of examples of self-X mechanisms in technical systems that in most cases are not based on biology:

- Fault-tolerant systems (from the 60s)
- Adaptive controllers (from the 70s)
- Metacognitive systems (90s)
- Autonomic Computing (00s)
- Adaptive service systems (00s)
- Organic Computing (10s)

All these systems observe themselves and use these observations to adapt a system's behavior to changing circumstances. These changes may be due to system-external disturbances or system-internal operational conditions. The adaptation to external changes has been widely investigated, but the adaptation to internal changes has received less attention.

In our own own case we investigate domain-neutral, application-neutral architectures for augmented autonomy based on model-based reflective adaptive controllers. Domain neutral means that we investigate architectures for any kind of system—e.g., mobile robots or chemical factories—and application neutral means that the architectures shall provide functionality for any kind of application—e.g., for system fault tolerance or dynamic service provision.

3.2 AN EXAMPLE: A METACONTROLLER FOR ROBOTS

Figure 3 shows an implementation of a self-aware system that improves resilience of a mobile robot.[10] The self-awareness mechanism is a metacontroller—a controller of a controller—that manages the operational state of the robot. This metacontroller has been designed to mitigate the resilience reduction due to potential faults in the control system of the robot.

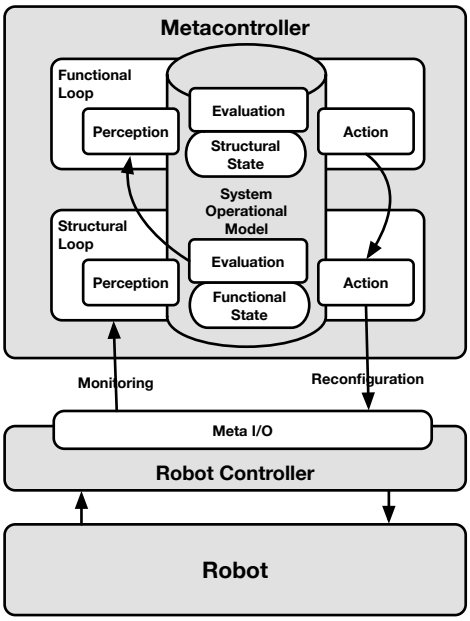


Figure 3. A metacontroller for robots developed for improving adaptivity of the robot controller.

The robot controller is a very complex distributed software system that can suffer transient or permanent faults in any of its components. The metacontroller monitorizes the state of the robot controller and acts upon it to keep system functionality by reorganizing its functional organization. This is similar to what humans do when overcoming some of the problems of becoming blind by learning to read with the fingers.

Function, functional state, and componential organization are core concepts in this approach. In this system this self-awareness mechanism provides reaction to disruption and improves mission-level resilience. And this is grounded in a self-model based on formally specified concepts concerning the system and its mission.

Figure 4 shows part of the formal ontology[11] that is used in the implementation of the perception, reasoning, and action mechanisms of the self-awareness engine. It enables the robot to reason about its own body and mind.

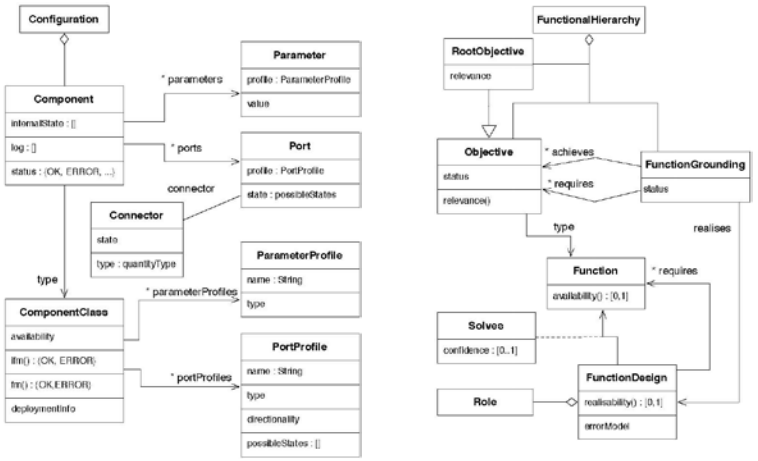


Figure 4. Part of the formal ontology that enables the robot to reason about its own body and mind.

3.3 INTO PHILOSOPHY

This research on machine self-awareness obviously enters philosophical waters.

The attempt to achieve engineering universality—any kind of system, any kind of environment, any kind of mission—implies that the *ontologies* and *architectural design patterns* that support the engineering processes must be general and not particular for the project at hand.

This requires a more scientific/philosophical approach to the conceptual problem that gets harder when human concepts enter into the picture (to provide an easy path for bioinspiration or to address knowledge and control integration in human-cyber systems).

Plenty of questions arise when we try to address self-awareness from a domain neutral perspective (i.e., a systems perspective that is based on concepts that are applicable both to humans and machines): What is Awareness? What is Self? What is Perception? What is Understanding? But these questions will be answered elsewhere because the rest of the article is dedicated to a problem in the science/philosophy of (machine) consciousness: **It is too centered in humans.**

This is somewhat understandable because humans are THE SPECIMENS of consciousness. But this is also a problem in general AI and Robotics, in convergent studies on philosophy of computers, and even in general cognitive science that is neglecting important results in intelligent systems engineering and losing opportunities for ground truth checks.

4. TOO MUCH ANTHROPOMORPHISM

Readers may have heard of Sophia; a humanoid robot who uses AI and human behavioral models to imitate human gestures and facial expressions. Sophia can maintain a simple conversation, answering simple questions. From an AI or robotics standpoint it is quite a low feat. Why does it get so much attention?

Raymond Tallis, in his book *The Explicit Animal*, offers us a clue in the form of a rant against functionalism:

> "Traditional mental contents disappear and the mind itself becomes an unremarkable and unspecial site in causal chains and computational procedures that begin before consciousness and extend be- yond it. Indeed, mind is scarcely a locus in its own right and certainly does not have its own space. It is a through road (or a small part of one) rather than a dwelling. Consciousness is voided of inwardness."[12]

For Tallis, causal theories of consciousness reduce mind to a set of input/output relations, with the net effect of effectively "emptying" consciousness. Causal links—the stuff machines are made with—seem not enough for the machinery of mind.

The same phenomenon can be found in any context where human mental traits and "similar" machine traits are put under the scope of the scientist or philosopher. For example, in a recent conference on philosophy of AI, a speaker raised the question "Is attention necessary for visual object classification?" A few slides later the speaker showed that Google was doing this without attention. So the answer of the question was NO. End. Surprisingly, the presentation continued with a long discussion about phenomena of human perception.

4.1 ANTHROPO-X

Cognitive Science and the Philosophy of Mind, and to some extent Artificial Intelligence and Robotics, are anthropocentric, anthroposcoped and anthropobiased. They focus on humans; they address mostly humans; they think of humans as special cases of mind and consciousness. Obviously, the human mind ranks quite high in the spectrum of mind. Maybe it is indeed the peak of the scale. But this does not qualify it as special in the same sense that elephants or whales are not special animals, however big.

However, the worst problem is that all these disciplines are also anthropomorphic: They shape all their theories using the human form. Protagoras seems still alive and man is used to measure all things mental.

This is not only wrong, but severely limiting. Anthropomorphism has very bad effects in consciousness research:

- Human consciousness traits are considered general consciousness traits. This has the consequence for artificial systems of posing extra, unneeded requirements for the implementation of cognitive engines for machines (see, for example, the wide literature on cognitive architectures).

- Some non-human traits are not properly addressed in the theories; because being out the the human spectrum are considered irrelevant concerning the achievement of machine consciousness.

4.2 RETHINKING CONSCIOUSNESS TRAITS

Some commonly accepted consciousness aspects shall be rethought under a non-chauvinistic light to achieve the generality that science and engineering require.

For example, consciousness *seriality* and *integration* have been hallmarks of some widely quoted theories of consciousness.[13] Functional departures from these are considered pathological, but this is only true under the anthropocentric perspective. Consciousness seriality implies that an agent can only have one stream of consciousness. However, from a general systems perspective, nothing prevents a machine from having several simultaneous streams.

This aspect of seriality is closely related to *attention*. According to Taylor,[14] attention is the crucial gateway to consciousness and architectural models of consciousness shall be based on attention mechanisms. However, using the same analysis as before, nothing prevents a machine from paying attention to several processes simultaneously. The rationale of attention mechanisms seems to be the efficient use of *limited* sensory processing resources (esp. at higher levels of the cognitive perception pipeline). But in the case of machines, if the machine architecture is scalable enough, it is in principle possible to incorporate perceptual resources as needed to pay concurrent attention to several processes. This is also related to the limited capacity trait of human consciousness.

Integration is another human trait that may be unnecessary in machines. Consciousness integration implies that the collection of experiences flowing up from the senses are integrated in a single experiential event. Dissociated experience is abnormal—pathological—in humans but it need not be so in machines. In fact, in some circumstances, being able to keep separated streams of consciousness may be a benefit for machines (e.g., for cloud-based services for conscious machines).

Concepts like subjectivity, individuality, consciousness ontogenesis and filogenesis, emotional experience, sensory qualia modalities, etc. all suffer under this same analysis. Any future, sound theory of consciousness shall necessarily deal with a wider spectrum of traits like bat audio qualia or robot LIDAR qualia.

5. CONCLUSIONS

Universality renders deep benefits. This has been widely demonstrated in science, for example, when the dynamics of falling objects on the earth surface was unified with the dynamics of celestial objects.

We must escape the trap of anthropomorphism to reach a suitable theory for artificial consciousness engineering. Just consider the history of "artificial flight," "artificial singing," or "artificial light." We need not create mini suns to illuminate our rooms. We don't need to copy, nor imitate, nor fake human consciousness for our machines. We don't need the whole iguana of mind; what we need are analysis and first principles.

Any general (non anthropocentric) consciousness research program will produce benefits also in the studies of human consciousness because it can provide inspiration for deeper, alternate views of human consciousness. For example, human brains have parallel activities (not serial but concurrent) in different levels of an heterarchy. Considerations coming from conscious distributed artificial systems will help clarify issues of individual minds—normal and pathological—and social consciousness.

Philosophy is a bold endeavor. It goes for the whole picture. Philosophy of mind shall be aware of this and realize that consciousness escapes humanity.

NOTES

1. Wasson,*System Engineering Analysis, Design, and Development: Concepts, Principles, and Practices*.
2. Sloman, "The Mind as a Control System"; Prescott et al., "Layered Control Architectures in Robots and Vertebrates"; Sterelny, *The Evolution of Agency and Other Essays*; Winning, "The Mechanistic and Normative Structure of Agency."
3. Sanz and Meystel, "Modeling, Self and Consciousness: Further Perspectives of AI Research."
4. Simon,*The Sciences of the Artificial*.
5. In fact, there are control systems engineering methods that do not perform this separation, addressing mind and body as a single whole by concurrent co-design or by embedding control forced dynamics by reengineering of the body.
6. Webb, "Animals Versus Animats: Or Why Not Model the Real Iguana?"
7. Russell and Norvig, *Artificial Intelligence: A Modern Approach*.
8. Sanz et al., "Consciousness, Meaning and the Future Phenomenology."
9. Conant and Ashby, "Every Good Regulator of a System Must Be a Model of That System"; M. Branicky et al., "A Unified Framework for Hybrid Control: Model and Optimal Control Theory."
10. Hernández et al., "A Self-Adaptation Framework Based on Functional Knowledge for Augmented Autonomy in Robots."
11. "Ontology" in knowledge engineering is a specification of a conceptualization. This specification is used to ground the use and interchange of information structures among humans and machines.
12. Tallis, *The Explicit Animal: A Defence of Human Consciousness*.
13. Tononi, "An Information Integration Theory of Consciousness."
14. Taylor, "An Attention-Based Control Model of Consciousness (CODAM)."

REFERENCES

Branicky, M., V. Borkar, and S. Mitter. "A Unified Framework for Hybrid Control: Model and Optimal Control Theory." *IEEE Transactions on Automatic Control* 43, no. 1 (1998): 31–45.

Conant, R. C., and W. R. Ashby. "Every Good Regulator of a System Must Be a Model of That System." *International Journal of Systems Science* 1, no. 2 (1970): 89–97.

Hernández, C., J. Bermejo-Alonso, and R. Sanz. "A Self-Adaptation Framework Based on Functional Knowledge for Augmented Autonomy in Robots." *Integrated Computer-Aided Engineering* 25 (2018): 157–72.

Prescott, T. J., P. Redgrave, and K. Gurney. "Layered Control Architectures in Robots and Vertebrates." *Adaptive Behavior* 7 (1999): 99–127.

Russell, S., and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 3rd ed., 2010.

Sanz, R., C. Hernández, and G. Sánchez. "Consciousness, Meaning and the Future Phenomenology." In Machine Consciousness: Self, Integration and Explanation, edited by R. Chrisley, R. Clowes, and S. Torrance, 55–60. York, UK, 2011.

Sanz, R., and A. Meystel. "Modeling, Self and Consciousness: Further Perspectives of AI Research." In *Proceedings of PerMIS '02, Performance Metrics for Intelligent Systems Workshop*. Gaithersburg, MD, USA, 2002.

Simon, H. A. *The Sciences of the Artificial*, 3rd edition. Cambridge, MA: The MIT Press, 1996.

Sloman, A. "The Mind as a Control System." In *Proceedings of the 1992 Royal Institute of Philosophy Conference on Philosophy and the Cognitive Sciences*, edited by C. Hookway and D. Peterson. Cambridge University Press, 1992.

Sterelny, K. *The Evolution of Agency and Other Essays*. Cambridge University Press, 2001.

Tallis, R. *The Explicit Animal: A Defence of Human Consciousness*. Palgrave Macmillan, 1999.

Taylor, J. G. "An Attention-Based Control Model of Consciousness (CODAM)." *Science and Consciousness Review*. (2002).

Tononi, G. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5, no. 42 (2004).

Wasson, C. S. *System Engineering Analysis, Design, and Development: Concepts, Principles, and Practices*, 2nd edition. Wiley Series in Systems Engineering and Management. Wiley, 2015.

Webb, B. "Animals Versus Animats: Or Why Not Model the Real Iguana?" *Adaptive Behavior* 17, no. 4 (2009): 269–86.

Winning, R. J. "The Mechanistic and Normative Structure of Agency." PhD thesis, University of California, San Diego, 2019.

PHYSICAL COMPUTATION: A MECHANISTIC ACCOUNT

Should Physical Computation Be Understood Mechanistically?

John Symons
UNIVERSITY OF KANSAS

Common sense tells us that a rock is not a computer but that your laptop is. While common sense is untroubled by the straightforward cases, it is of little help with the exceptions and the exotic cases: Does a broken or faulty laptop count as a computer? And what if aliens used devices that looked like rocks to compute in ways we are unable to understand? Is it sometimes correct to think of rivers, forests, or strands of DNA as computers? Is the central nervous system of an animal a computer? Some people like to think of brains and minds as computers, but isn't this an odd claim given how unlike ordinary computers biological systems seem to be? We could begin to answer such questions in a principled way if we had a plausible theoretical account of the kinds of things that are physical computers. Currently we don't have such an account.

Our uncertainty with respect to the nature of physical computers contrasts sharply with our excellent understanding of the mathematics of computation; we have a well-articulated theory of computable functions, and we have a good mathematical model of systems that compute. The formalism of computability is elegant, clear, and relatively easy to understand. We say that a system that computes is a finite state automaton. A finite state automaton is a *mathematical object* that can be in one state at a time and that switches between states in a rule-governed way. We can reason clearly about the properties of such mathematical objects and have discovered some of their limits. Where things become less clear is when we ask what it means to say that some *physical object* computes.

One response is to simply dismiss questions about physical computation as being nothing more than a matter of interpretation. Arguably, one can construct an interpretation whereby *any* physical object or any arbitrary mereological sum of objects (your thumb, a lake, Jupiter, or the pile of crumbs on the counter) can be interpreted as being *any* finite-state automaton. Hilary Putnam defended a position roughly along these lines.[1] However, as a response to the problem of distinguishing physical computers from noncomputers, this strategy is philosophically unsatisfying and scientifically unhelpful.[2]

Stating an adequate criterion for distinguishing physical computers from noncomputers has proven difficult for philosophers. Clearly, the problem is not solved simply in virtue of having a good mathematical theory of computability. More is needed. This is because a satisfactory account of physical computation, unlike a mathematical theory of computability, should provide individuation conditions that distinguish objects and processes that compute from those that do not. In his recent book *Physical Computation: A Mechanistic Approach*, Gualtiero Piccinini defends a novel and appealing theory of physical computation. On his view, a physical computer is "a mechanism whose teleological function is to perform a physical computation and a physical computation is the manipulation of a medium-independent vehicle according to a rule."[3] In this paper, I will examine the strengths and limitations of this position, concentrating on the question of whether the mechanistic approach has the resources to provide a satisfying account of the individuation of physical computers. Unfortunately, the concept of mechanism cannot provide an illuminating account of the metaphysics of physical computation. This is because the notion of mechanism is insufficiently fundamental and insufficiently general. Nevertheless, Piccinini's book provides an accurate map of the philosophical problems associated with the individuation of physical computers and is filled with important distinctions and insights. Piccinini offers the best attempt to date to answer these questions.

Piccinini shares a view of mechanism with the New Mechanists in philosophy of biology and philosophy of cognitive science. The New Mechanist approach began to take shape in the early 1990s with the work of Bill Bechtel and Bob Richardson, especially in their book *Discovering Complexity*. Today, the canonical reference for the New Mechanist position is Peter Machamer, Lindley Darden, and Carl Craver's "Thinking about Mechanisms." In his *Stanford Encyclopedia of Philosophy* entry "Mechanism in Science," Carl Craver writes that all definitions of mechanism involve four characteristic features: (1) a phenomenon to be explained, (2) parts, (3) causings, and (4) organization.[4] According to Bechtel and Abrahamsen, for example, "[a] mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena."[5]

Very broadly speaking, the New Mechanists provide an account of mechanism that they have gleaned from the manner in which some biologists and some neuroscientists give explanations. They notice that in many fields in the biological sciences explanations seem to work by showing how structured parts and processes are orchestrated so as to be responsible for the way things appear to happen. What is meant by orchestration and responsibility is left relatively loosely characterized. Nevertheless, New Mechanists correctly point to the ubiquity and wide acceptance of such patterns of explanation in biology. On this view, the job of at least some biologists is to discover the mechanisms that produce some phenomena.

The New Mechanists are a group of pragmatically inclined philosophers who are guided by what they see as the explanatory norms and practices of the scientific communities they know best. In this spirit, the New Mechanists sometimes seem to imply that their accounts capture a metaphysically agnostic core of ordinary explanatory practice in the sciences. While New Mechanists assume a modest posture in relation to scientific practice, there is (or should be) more to New Mechanism than just a neutral way of parsing accepted styles of explanation in scientific communities. Mechanists regard the explanatory practices of scientists as indicating that there really are distinguishable parts of mechanisms with well-defined relations and activities or operations. In this sense, New Mechanism has non-trivial metaphysical commitments.

Piccinini puts this view of mechanism into action in his description of physical computers. On his view, physical computers can be understood in terms of their component parts, their *functions*, and their organization. Physical computers are objects whose function, according to Piccinini, is to perform physical computations. Physical computation is the manipulation of medium-independent vehicles according to rules. Rules and functions resist reduction to mechanism, but on Piccinini's account the physical systems that follow those rules are ultimately nothing more than mechanisms.

In practice, physical computers are certainly picked out by reference to their functions, and those functions are determined (at least in part) by reference to abstract rules. However, a description of what the physical computer *is* (as opposed to how we happen to identify it) need not include mention of those rules and functions. The role of rules and functions in his presentation is intended to serve as a way of distinguishing the subset of mechanisms that count as physical computers (namely, those that serve the function of following rules) from those that do not. Thus, Piccinini does not claim that we rely exclusively on the notion of mechanism in the process of *identifying* some objects as physical computers. In fact, it might be a matter of epistemic necessity that our identification of physical computers depends on reference to rules and functions. Nevertheless, the way we happen to distinguish physical computers from other mechanisms is not ultimately relevant to what they are. Identification is not the same as individuation.

On Piccinini's account, mechanism plays a role in helping us to understand the metaphysics of physical computation. One important reason to give an account of physical computation in terms of mechanism is to sidestep Putnam's concerns about the challenge of determining a unique mapping from abstract computational characterizations to physical implementation. Putnam argued that every ordinary open physical system can be interpreted as implementing every finite-state automaton. This means that every physical object can justifiably be interpreted as a computer,[6] but it would also render the mapping account useless as a means of individuating physical computers. Piccinini claims that the mapping approach poses the problem incorrectly and so he does not answer Putnam's challenge. His strategy is to abandon the question of how we should interpret the mapping between states of a physical system and states of an abstract computer. He suggests that even amended or strengthened mapping accounts will trivialize the claim that a physical system computes.[7] While Piccinini does not employ the distinction between individuation and identification in the manner discussed above, the mapping approach can be understood as addressing epistemic considerations involved in identifying physical computers whereas the correct approach would address the problem of individuating physical computers. In this sense, the mechanistic account can be understood as offering an alternative strategy for thinking about those individuation conditions. The activity of physical computation is a purely mechanistic matter on Piccinini's view.

This explanatory strategy allows him to claim that what it is to *be* a physical computer does not depend on representational or semantic concepts. If correct, this would mark an important step forward insofar as representation and semantics have been central to many previous accounts of computation and have presented deep conceptual difficulties familiar to the traditions of philosophy of mind and philosophy of language. Central to Piccinini's contention that the mechanistic approach can do without representation and semantics is his understanding of how physical computers perform *concrete computations*. "A physical system is a computing system," he writes, "iff it's a system that performs concrete computations."[8] Concrete computations support the teleological function of the physical computer insofar as the rules characterizing the transformation of vehicles are fixed by the purpose of the machine. Nevertheless, that transformation itself can be described independently of abstract rules. Rather than individuating these vehicles by reference to their semantical roles, the non-semantic properties of strings of discrete states that figure in program-controlled computers can serve as the basis of the explanation of the operation of those systems. "Different bits and pieces of these strings of states have different effects on the machine."[9] Because of this, he argues, we can describe the operation of sub-strings and states without mentioning what their semantics are. Moreover, according to Piccinini, the mechanistic account of vehicles can explain syntactical properties of digital computation. Since mechanism is a more basic notion than the syntax of a language, Piccinini can argue that a mechanistic approach has the additional advantage of providing an account of non-digital forms of computation whereas the syntactical approach fails to do so. The promise of the mechanistic approach is that only non-semantic, non-representational, and even non-syntactic individuation

conditions for the vehicles of computation will figure in mechanistic explanations.

The mechanistic approach seems more metaphysically basic than, for example, the semantic or representational level of analysis. However, mechanistic approaches have a hard time shedding light on the problem of individuation.[10] To begin with, the metaphysical significance of mechanism is not clear. In part, this is because the mechanistic approach is primarily derived from reflections on scientific explanation as a practice rather than on those aspects of reality that ground the reliability of scientific explanations. It is obviously true that in many areas of biology explanations are given in mechanistic terms. But this is not an explanation of why mechanistic accounts are widely accepted in those areas or why they have been so successful. If anything, the success of mechanistic explanation in these areas itself demands explanation. Worth mentioning too is the need for an account of the notion of part, cause, and organization that undergird the notion of mechanism itself.

A second reason for the difficulty of using mechanistic accounts to shed light on individuation and other metaphysical topics arises from the fact that mechanism is both an insufficiently general and an insufficiently fundamental notion. With respect to generality, notice, for example, that the mereological commitments of New Mechanism make it inapplicable to some kinds of objects and phenomena. The New Mechanists acknowledge that their view is only applicable to some regions of even more mundane domains of scientific explanation. Even in biology, as Skipper and Millstein point out, and as Craver and Tabery acknowledge, the mechanistic notion of part is difficult to sustain both at the level of the very small in well-understood biochemical phenomena and at the level of the very large in natural selection.[11] Our best understanding of nature does not support the idea that everything can be understood or explained in terms defended by the New Mechanists; therefore, the mechanistic approach is unlikely to have the resources to answer general questions concerning individuation.

With respect to fundamentality, it is difficult to reconcile New Mechanism with basic physics. For example, it is difficult to understand field-theoretic explanations within a mechanistic framework. Furthermore, the ontological implications of quantum mechanics pose a challenge insofar as New Mechanism seems to rely on local causal interactions and the idea of isolable parts with definite properties.

In the case of physical computation, there is a sense in which such metaphysical concerns might seem irrelevant. Most physical computers are artifacts with very special characteristics. The purposes of most physical computers are dictated by the demands of the target markets of their manufacturers. Since the practice of manufacturers serves specific commercial or scientific purposes and since most computers are built from pre-fabricated parts according to a plan for organization, it is appropriate to characterize these physical computers as functional mechanisms.

The component parts of physical computers also have parts, of course. However, the mereological ground floor for individuating physical computers (on Piccinini's view) is fixed by the logic of software.[12] At bottom, the primitive *computationally relevant* components for conventional computers are logic gates. In principle, any bistable system can serve as a primitive component for a computing technology if it allows us to reliably manipulate whether that system is in one or the other equilibrium state. There is nothing relevant to computational function *per se* at more fundamental levels than the level of such bistable systems.

The mechanistic approach to physical computation does not tell us why or how the primitive components of a conventional physical computer work as they do. As far as the mechanistic perspective is concerned, once one has reliably manipulable bistable systems to build on, everything underneath can be ignored. However, the promise of the mechanistic approach to physical computation had been the possibility of providing an explanation of the relevant primitive components: "The mechanistic explanation of a primitive computing component—say, an AND gate—explains how that component exhibits its specific input-output behavior. In our example, the components of a particular electrical circuit, their properties, and their configuration explains how it realizes an AND gate."[13] Unfortunately, even in relatively mundane engineering contexts, for example, in the explanation of modern transistors and circuits, New Mechanist-style explanations fall short. The New Mechanist approach will not get very far, for example, in the understanding of semi-conductors or in the understanding of field effect transistors more generally. The details of how fields behave are completely opaque if one's only explanatory resource is mechanism. Mechanisms are composed of objects or processes with definite properties. The quantum mechanical account of the behavior of fields does not assume local causal interactions of the mechanist kind nor does it assume objects with definite properties.[14] Mechanists might reject this concern given that once reliable bistable systems emerge at some level everything underneath can be ignored. However, the fundamental physical nature of these bistable systems turns out to be relevant to their operation and must be taken into account in engineering contexts. Consider, for instance, the consequences of miniaturization. As the miniaturization of transistors continues, quantum tunneling will make it difficult to insulate the relevant parts of the circuits. In order to build the primitive bistable components of computers, quantum-level behaviors would be unavoidable. In this context one solution that has been proposed involves exploiting quantum tunneling as part of the operation of the circuits themselves. There will be no explanation of such components and no light will be shed on practical solutions from a mechanistic perspective.

These challenges and details fall below the level of the primitive components of the mechanistic account of computation. The trouble here is that these primitive components are picked out by the logic of a particular kind of software. Thus the mereological ground floor for the mechanist's treatment of computation is established by reference to the abstract characteristics of computation rather than being explained by, or grounded in, mechanism. If we were concerned about questions like, What are the

physical constraints on computation? How do bistable systems emerge in physical reality? or any number of other questions about the physical instantiation of computation, the mechanistic approach will not satisfy.

ACKNOWLEDGMENTS

Many thanks to Mark Bickhard, Piotr Boltuc, Frances Egan, Jack Horner, Corey Maley, Marcin Miłkowski, Nico Orlandi, Gualtiero Piccinini, and Sarah Robins for helpful discussions of this material.

This work is supported by The National Security Agency through the Science of Security initiative contract #H98230-18-D-0009.

NOTES

1. See especially Putnam, *Representation and Reality*.
2. Buechner,*Gödel, Putnam, and Functionalism: A New Reading of 'Representation and Reality'*.
3. Piccinini,*Physical Computation: A Mechanistic Account*, 10.
4. Cravor and Tabery, "Mechanisms in Science."
5. Bechtel and Abrahamsen, "Explanation: A Mechanist Alternative," 523.
6. Putnam,*Representation and Reality*, 120–25.
7. Piccinini,*Physical Computation: A Mechanistic Account*, 22.
8. Ibid., 118.
9. Ibid., 45.
10. Symons, "The Individuality of Artifacts and Organisms."
11. Skipper and Millstein, "Thinking about Evolutionary Mechanisms: Natural Selection"; Craver and Tabery, "Mechanisms in Science."
12. Horner and Symons, "Understanding Error Rates in Software Engineering: Conceptual, Empirical, and Experimental Approaches."
13. Ibid., 155.
14. See Kuhlmann and Glennan, "On the Relation Between Quantum Mechanical and Neo-Mechanistic Ontologies and Explanatory Strategies," 338.

REFERENCES

Bechtel, William, and Adele Abrahamsen. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36, no. 2 (2005): 421–41.

Bechtel, William, and Robert C. Richardson. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. MIT Press, 1993.

Buechner, Jeff. *Gödel, Putnam, and Functionalism: A New Reading of 'Representation and Reality'*. MIT Press, 2008.

Craver, Carl, and Tabery, James, "Mechanisms in Science." *The Stanford Encyclopedia of Philosophy*. Spring 2017 Edition. Edward N. Zalta (ed.). Available at https://plato.stanford.edu/archives/spr2017/entries/science-mechanisms/. Last Accessed January 2, 2019.

Horner, Jack, and John Symons. "Understanding Error Rates in Software Engineering: Conceptual, Empirical, and Experimental Approaches." *Philosophy & Technology* (2019): 1–16.

Kuhlmann, Meinard, and Stuart Glennan. "On the Relation Between Quantum Mechanical and Neo-Mechanistic Ontologies and Explanatory Strategies." *European Journal for Philosophy of Science* 4, no. 3 (2014): 337–59.

Machamer, Peter, Lindley Darden, and Carl F. Craver. "Thinking about Mechanisms." *Philosophy of science* 67, no. 1 (2000): 1–25.

Piccinini, Gualtiero. *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press, 2015.

Putnam, Hilary. *Representation and Reality*. MIT Press, 1988.

Skipper Jr., Robert A., and Roberta L. Millstein. "Thinking about Evolutionary Mechanisms: Natural Selection." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36, no. 2 (2005): 327–47.

Symons, John. "The Individuality of Artifacts and Organisms." *History and Philosophy of the Life Sciences* (2010): 233–46.

Commentary on Gualtiero Piccinini's Physical Computation: A Mechanistic Perspective

Martin Roth

DRAKE UNIVERSITY

Gualtiero Piccinini's *Physical Computation: A Mechanistic Perspective* is an example of first-rate scholarship—rigorous, clear, well-informed, and well-argued—and though I have some criticisms of the perspective Piccinini develops in his book, I suspect that what emerges is not so much a fundamental disagreement as it is a difference in emphasis. My criticisms focus on themes developed in Chapter 5—"From Functional Analysis to Mechanistic Explanation." In that chapter Piccinini challenges the "received view" according to which functional analysis is distinct and autonomous from mechanistic explanation. As Piccinini acknowledges, the views developed in that chapter draw heavily from a 2011 paper that Piccinini wrote with Carl Craver.[1] My criticisms will be directed at that paper.[2]

When we identify something functionally—a mousetrap, a gene, a legislature—we identify it in terms of what it does. Many biological terms have both a functional and an anatomical sense: an artificial heart is a heart by function but is not an anatomical heart, and computational neuroscience was conceived when the word "brain" became a functional term as well as an anatomical one. Functional analysis is the attempt to explain the properties of complex systems—especially their characteristic capacities—by the analysis of a systemic property into organized interaction among other simpler systemic properties or properties of component subsystems. This explanation-by-analysis is *functional* analysis because it identifies analyzing properties in terms of what they do or contribute, rather than in terms of their intrinsic constitutions. For example, a circuit diagram describes or specifies a circuit in a way that abstracts away from how the components, including the "wires," are actually made. The strategy of explaining the capacities of complex systems by functional analysis is ubiquitous in science and engineering, and by no means special to psychology.

From the point of view of functional analysis, capacities are dispositional properties, and the dispositional properties of a complex system are explained by exhibiting their manifestations as the disciplined manifestation of dispositions that are components of the target disposition, or by the disciplined interaction of the dispositions of the system's component parts. It should be obvious that the explanatory targets of this sort of analysis are not points in state space or particular trajectories through it. Rather, the aim of this kind of analysis is to appeal to a system's design in order to explain why one finds the trajectories one does and not others. The design provides a model of

the state space and constrains the possible paths through it, thereby explaining certain regularities in the system's behavior. More generally, the strategy is to explain the capacities of a complex system by exhibiting the abstract functional design of that system—to show, in short, that a system with a certain design is bound to have the capacity in question. Designs can do this because functional terms pick out the causal powers that are relevant to the capacity being analyzed. Functional terms are in this sense *causal relevance filters*: by selecting from the myriad causal consequences of a system's states, processes, or mechanisms those that are relevant to the target capacity, functional characterization makes the contributions of those states, processes, or mechanisms transparent. It is precisely this transparency that enables us to understand *why* anything that possesses these states, processes, or mechanisms is bound to have the capacity in question. Without this filtering, we are simply left with a welter of noisy detail with no indication of what is relevant and what is a mere by-product of this or that implementation. Causal relevance filtering is, therefore, just abstraction from the implementation details that are irrelevant to the achievement of the targeted capacity. Implementations that differ in those details but retain the design will thus all exhibit the targeted capacity. In this way, the possibility of multiple realization is an inevitable consequence of causal relevance filtering, and so it should come as no surprise to find that functional analyses subsume causal paths that have heterogeneous implementations.

However, it would be a mistake to wed the explanatory power of functional analysis to assumptions about actual multiple realization, for even if there is only one nomologically possible way to implement a design, giving implementation details that go beyond what is specified by an analysis adds nothing to the explanation provided by the design. For example, suppose there is just one nomologically possible way to implement a doorstop—say, by being a particular configuration of rubber. In this case, it would be plausible to hold that being a doorstop—the type—is identical to being a particular configuration of rubber—the type. Because type-type identities give you property reductions, being a doorstop would thus reduce to being a particular configuration of rubber. But a functional analysis that specifies something *as a doorstop* would still be autonomous, in the following sense. Being a particular configuration of rubber comes with any number of causal powers. One of those powers is stopping doors, and in the context of the imagined functional analysis, stopping doors is the only causal power of this particular configuration of rubber that matters to having the target capacity. If in our analysis we replace the word "doorstop" with the phase "rubber configured thus and so," we won't lose anything as far as the causation goes. However, we will lose the transparency functional analysis affords *unless we specify explicitly that stopping doors is the relevant causal power*. But then the explanation is tantamount to the explanation given in terms of the word "doorstop," i.e., the explanation does not give us anything beyond what is provided by the functional analysis itself.

If we focus on the causal explanation of events and assume type-type identity, then framing explanations in terms of the word "doorstop" is guaranteed to give you nothing beyond what framing explanations in terms of the phrase "rubber configured thus and so" gives you, and this is why it has been generally assumed that reduction is incompatible with autonomy. From the perspective of functional analysis, by contrast, autonomy can live with reduction. Design explanations are autonomous in the sense that they do not require "completion" by annexing implementation details, e.g., in the case imagined above, it is irrelevant to explaining the target capacity whether a specific doorstop is a particular configuration of rubber. But design explanations are also autonomous in the sense that adding implementation details would *undermine* the transparency provided by causal relevance filtering and thereby obviate the understanding provided by the design. A doorstop may *be* a particular configuration of rubber, but replacing the word "doorstop" with the phrase "rubber configured thus and so" masks the information needed to understand why a system has the target capacity.

I am sympathetic to the thought that complete knowledge of implementation details would contribute to a fuller understanding of those systems whose capacities are targeted by functional analysis. Indeed, such details are necessary for understanding how a system manages to have the very causal powers that are picked out by functional analysis. But having a fuller understanding of a system, in this sense, is not the same thing as having a more complete explanation of the capacity targeted for functional analysis. For example, when we analyze the capacity to multiply numbers in terms of a partial products algorithm, the specification of the algorithm tells us nothing about the states, processes, or mechanisms of a system that implements the algorithm (except in the trivial sense that the states, processes, or mechanisms of any system that implements the algorithm are sufficient for implementing it). However, as far explaining the capacity goes—what we might call the "multiplication effect"—the analysis provided by the algorithm is complete, i.e., the analysis allows us to understand why any system that has the capacity for computing the algorithm *ipso facto* exhibits the multiplication effect. Because details about how the algorithm is implemented add nothing to the analysis, such details are irrelevant to the explanation of the capacity.

The perspective I have outlined here suggests that we need to distinguish two kinds of explanations, what I call *horizontal* and *vertical* explanations. Horizontal explanations explain capacities by appeal to a design that is specified by functional analysis. They answer the question "Why does system S have capacity C?" by specifying some design D. Vertical explanations specify implementations. They answer the question "How is design D realized in system S?" Neither type of explanation is subsumption under law. And neither is in the business of explaining individual events. The explananda are, respectively, capacities and designs.

I suspect that the tendency to conflate *explaining a capacity via functional analysis* with *explaining how a functional analysis (a design) is implemented* has led to a misunderstanding concerning the relationship between functional analysis and mechanistic explanation. Following

Bechtel and Abrahamsen, a mechanism "is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena."[3] As I see it, the goal of discovering and specifying mechanisms is often or largely undertaken to explain how the analyzing capacities specified by a functional analysis are implemented in some system. In this way, though the horizontal explanations provided by functional analysis are autonomous from the vertical explanations provided by specifying mechanisms, the two explanations complement each other.

However, Piccinini and Craver appear to challenge this claim of autonomy.[4] They argue that functional analyses are "mechanism sketches": functional analyses and the design explanations they provide are "incomplete" until filled out with implementation details, and in that way, the explanations provided by functional analysis are not autonomous. However, I think their argument involves a misidentification of the relevant explanatory targets of functional analysis—capacities—and a correlative conflation of explanation and confirmation. I'll take these up in turn. Piccinini and Craver write that "Descriptions of mechanisms . . . can be more or less complete. Incomplete models—with gaps, question-marks, filler-terms, or hand-waving boxes and arrows—are mechanism sketches. Mechanism sketches are incomplete because they leave out crucial details about how the mechanism works."[5] Sketches being what they are, I have no quarrel with the claim that mechanism sketches are incomplete, and insofar as mechanistic explanations explain by showing how a mechanism works, I agree that filling in the missing details of a mechanism sketch can lead to a more complete mechanistic explanation. The crucial issue here, however, is whether functional analyses should be viewed as mechanism sketches. To motivate the claim that functional analyses are mechanism sketches, we have to assume that abstraction from implementation detail inevitably leaves out something crucial to the *analytical explanation of a target capacity*, something that implementation details would provide. But as I've already argued, the opposite is in fact true; adding implementation details obfuscates the understanding provided by functional analysis.

Instead of favoring the autonomy of functional analysis, however, Piccinini and Craver think that abstraction from implementation details actually works against claims of autonomy. They write:

> Autonomist psychology—the search for functional analysis without direct constraints from neural structures—usually goes hand in hand with the assumption that each psychological capacity has a unique functional decomposition (which in turn may have multiple realizers). But there is evidence that...several functional decompositions may all be correct across different species, different members of the same species, and even different time-slices of an individual organism. Yet the typical outcome of autonomist psychology is a single functional analysis of a given capacity. Even assuming for the sake of the argument that autonomist psychology stumbles on one among the correct functional analyses, autonomist psychology is bound to miss the other functional analyses that are also correct. The way around this problem is to let functional analysis be constrained by neural structures—that is, to abandon autonomist psychology in favor of integrating psychology and neuroscience.[6]

I think this argument conflates *explanatory* autonomy with *confirmational* autonomy. If a capacity admits of more than one analysis, merely providing *an* analysis will, of course, leave open the question of whether the analysis provided correctly describes how a system manages to have the capacity in question (assuming it does have the capacity). Knowledge of neural structures is undoubtedly relevant to settling the question of which analysis is correct, but bringing such knowledge to bear in this instance would be an exercise in confirming a proposed analysis, not explaining a capacity. Suppose there are two possible analyses, A and B, for some capacity C, and the neurological data suggests that analysis A is implemented in system S. The explanation of capacity C in S is provided *by A*, not by the neural structures evidence about which confirms A.

Arguably, nothing enjoys confirmational autonomy from anything else. As such, neuroscience that makes well-confirmed psychological capacities impossible or unlikely needs revision as much as a design hypothesis in psychology that appears to have no plausible neural implementation. I thus agree with Piccinini and Craver's claim that "psychologists ought to let knowledge of neural mechanisms constrain their hypotheses just like neuroscientists ought to let knowledge of psychological functions constrain theirs."[7] This is an invitation to those working in psychology departments and neuroscience departments to talk to each other, and if this is enough to show that psychology is not autonomous from neuroscience, then so much the worse for the autonomy of psychology. But defending the autonomy of *functional analysis* is not the same thing as defending the autonomy of *psychology*. Functional analysis is an explanatory strategy, not a scientific discipline, and when we are careful to distinguish horizontal and vertical explanations, and distinguish confirmation and explanation, the autonomy of functional analysis emerges as unproblematic.

NOTES

1. G. Piccinini and C. Craver. "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese* 183, no. 3 (2011): 283–311.
2. My decision to comment on that paper stems largely from the fact that Robert Cummins and I have criticized that paper in another work ("Neuroscience, Psychology, Reduction, and Functional Analysis," in *Explanation and Integration in Mind and Brain Science*, ed. D. Kaplan [Oxford University Press: Oxford. 2018]), and I am interested in how Piccinini responds to those criticisms. What follows draws heavily from the aforementioned paper.
3. W. Bechtel and A. Abrahamsen. "Phenomena and Mechanisms: Putting the Symbolic, Connectionist, and Dynamical Systems Debate in Broader Perspective," in *Contemporary Debates in Cognitive Science*, ed. R. Stainton (Blackwell: Malden, MA, 2006), 162.
4. Piccinini and Craver, "Integrating Psychology and Neuroscience."
5. Ibid., 292.

6. Ibid., 285.
7. Ibid.

Defending the Mapping Account of Physical Computation

Frances Egan
RUTGERS UNIVERSITY

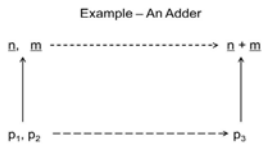
I will defend a version of the *mapping account* of computation, an account that Piccinini rejects in chapter two of *Physical Computation* in favor, ultimately, of a mechanistic view.[1] I will argue that a characterization of the sort specified by the mapping account is required even if one endorses a mechanistic view of computation. In fact, the two views should not be seen as competitors; both are required for a full account of a computing mechanism.

According to the version of the mapping account that I favor, a physical system S computes a function F just in case:

i. There is a mapping from (equivalence classes of) physical states of S to the arguments and values of F, such that:

ii. Causal state-transitions between the physical states mirror the formal dependency relations between the arguments and values of F.

An example may be helpful:



A physical system computes the addition function just in case there exists a mapping from physical state types to numbers, such that physical state types related by a causal state-transition relation ((p1, p2)→p3) are mapped to numbers n, m, and n+m related as addends and sums. Whenever the system goes into the physical state specified under the mapping as n, and goes into the physical state specified under the mapping as m, it is caused to go into the physical state specified under the mapping as n+m.

The mapping specifies, for a given physical system, *what* the system computes, and *how* it computes it—by transitioning through a sequence of causally related physical states. In other words, the mapping characterizes, at a high level of abstraction, the relevant causal organization in virtue of which the physical system is able to compute the specified function. A mapping of this sort has to be true if the physical system in fact computes the function. But a more perspicuous account of the mechanism's causal organization may characterize it in terms of structures, processes, and algorithms—in other words, in terms of components and their activities, as mechanists would insist.[2] The causal dependencies specified at the lower level of the mapping (the causal state transitions) will be consequences of this organization. The mapping isn't intended to be, and doesn't replace, an explanatory theory of a computing mechanism, that is, a perspicuous explanation of how the mechanism works.

According to Piccinini's mechanistic view of physical computation:

A physical system is a computing system just in case it has the following characteristics:

- It is a functional mechanism.
- One of its functions is to manipulate vehicles based solely on differences between different portions of the vehicles according to a rule defined over the vehicles. (275)

Let me make a couple of brief remarks about Piccinini's mechanistic account. Computers, both natural and artifactual, are undoubtedly functional mechanisms. They typically have specific teleological functions—to compute arithmetical functions, to detect the 3D structure of the scene, to enable navigation, to parse incoming speech—and more generally (and in the case of modern digital computers), to execute cognitive tasks. I find it implausible that computers have the function cited in the mechanistic account. Rather, manipulating vehicles based on differences in their parts according to a rule is the common *means* that computing mechanisms deploy to achieve their specific functions (parsing input strings, adding, and so on). Consider an analogy: applying upward pressure on a cork is a common means that various kinds of corkscrews deploy to achieve their function of removing corks from bottles. But I am not inclined to press this point too hard because I am not sure how the issue would be settled. The question of a natural computer's function would normally be settled by appeal to its manifest cognitive capacities (navigation, parsing speech, etc.); for artifacts, by appeal to the cognitive task the computer has been designed to achieve. In any event, the mechanistic account is missing a key component—specifying *what* function (in the mathematical sense, now) the system is computing when it manipulates vehicles in accordance with a rule; in other words, it is missing precisely what the mapping account provides.

I turn now to consider how well the mapping account satisfies Piccinini's desiderata for an adequate account of physical computation. I will focus on the two that might be supposed to cause the most trouble.

Desideratum (4): *The wrong things don't compute* – the account shouldn't entail that obvious non-computers in fact compute. Famously, this is where mapping accounts supposedly fail—they are supposedly *too liberal*. Putnam and Searle have argued that rocks, walls, and indeed all physical systems turn out to be computers because their physical states can be mapped to computational

descriptions. If computational descriptions apply to everything then computational "explanations" are trivial.

Here's what I think a mapping theorist should say in response to the triviality objection. First, the requirement that the physical states specified at the lower level of the mapping be *causally related* will eliminate many spurious mappings. Suppose, then, that we simply accept that for a physical system to compute a function just is for it to have the causal organization specified by the mapping. What would the possibility of unintended realizations imply for explanatory practice in cognitive science? In particular, would it make computational explanation trivial? In a word, *no*. To see why not, we need to focus on the *context* in which computational explanations are typically offered. Theorists deploy computational descriptions to explain the *manifest* cognitive capacities of physical systems such as ourselves, capacities such as adding, understanding, and producing speech, seeing the three-dimensional layout of the scene, and so on. (Rocks and walls have no manifest cognitive capacities.) With the target capacity in mind, the theorist hypothesizes that the system computes some well-defined function (in the mathematical sense), and spells out how computing this function would explain the system's observed success at the cognitive task. Justifying the computational description requires explaining how computing the value of the function contributes to the exercise of the cognitive capacity. For example, in David Marr's (1982) account of early vision, computing the Laplacian of a Gaussian of the retinal array produces a smoothed output that facilitates the detection of sharp discontinuities in intensity gradients across the retina, and hence the detection of significant boundaries in the scene. In other words, the computational description is justified by reference to the use to which the computation is put in the exercise of a manifest cognitive capacity. According to the mapping account, the theorist is thereby committed to the system actually having the causal organization specified by the lower level of the mapping. That is hardly trivial. That there may be other systems with this same causal organization, which, *if they were suitably hooked up to other internal mechanisms and situated in the right environment* (and these, it should be emphasized, are substantive constraints!) would enable the larger system to achieve the cognitive task, is simply not relevant.[3]

Desideratum (5): *The account should explain miscomputation*, i.e., "it should explain what it means for a concrete computation to give the *wrong* output." (24) I am not sure why Piccinini thinks the mapping account can't do this; in fact, it is one of the advantages of the view that an account of miscomputation falls right out of it.

Let's return to the addition example. When the system goes into the physical state interpreted under the mapping as 57 and goes into the physical state interpreted under the mapping as 43, and then is caused to go into the physical state interpreted under the mapping as 100, it correctly adds. If instead it were to go into the physical state interpreted under the mapping as 99 it would *miscompute* or *make a mistake*. On the other hand, if its battery died or it was dropped into the bathtub it would go into a physical state not interpreted under the mapping at all; it would not miscompute, it would *malfunction*. (To preserve the way we normally talk, a miscomputation may be seen as a special kind of malfunction.) So the specification of the function computed (here, addition) and the interpretation (here, the mapping of physical states to addends and sums) provides all we need to partition the behavioral space into (i) correct computations, (ii) miscalculations or mistakes, and (iii) malfunctions (states that receive no interpretation under the mapping). The specification of the function computed and the interpretation together provide the *norm* necessary to underwrite this three-fold distinction.

In summary, I think I have assuaged what Piccinini takes to be the most serious worries with the mapping account.

NOTES

1. For versions of mapping accounts, see Cummins, *Meaning and Mental Representation*, and Chalmers, "A Computational Foundation for the Study of Cognition," among others.
2. For other mechanistic accounts of physical computation, see Fresco, *Physical Computation and Cognitive Science*, and Milkowski, *Explaining the Computational Mind*.
3. See Egan, "Metaphysics and Computational Cognitive Science: Let's Not Let the Tail Wag the Dog," for elaboration of this argument.

REFERENCES

Chalmers, D. J. "A Computational Foundation for the Study of Cognition." *Journal of Cognitive Science* 12 (2011): 323–57.

Cummins, R. *Meaning and Mental Representation*. Cambridge, MA: MIT Press, 1989.

Egan, F. "Metaphysics and Computational Cognitive Science: Let's Not Let the Tail Wag the Dog." *The Journal of Cognitive Science* 13 (2012): 39–49.

Fresco, N. *Physical Computation and Cognitive Science*. New York: Springer, 2013.

Marr, D. *Vision*, New York: Freeman, 1982.

Milkowski, M. *Explaining the Computational Mind*. Cambridge, MA: MIT Press, 2013.

Comments on Gualtiero Piccinini, Physical Computation: A Mechanistic Account

Nico Orlandi
UNIVERSITY OF CALIFORNIA, SANTA CRUZ

Gualtiero Piccinini's new book is an insightful, informative, and admirably clear project. The book aims to give a plausible account of what concrete computation consists in. The account meets certain *desiderata*, and it contrasts in some important respects both with the orthodoxy that regards computation and representation as going together, and with the orthodoxy that regards computation as just a matter of mapping. Part of the problem with the two orthodoxies, Piccinini argues, is that they propose, respectively, a too restrictive and an overly liberal understanding of computation. Piccinini's book aims to delineate a middle passage where just the right things compute and the rest don't.

The centerpiece idea of the book is the *mechanistic account of Computation*. In these comments, I focus on getting clearer on the account. Since the comments must be brief, I focus on two questions in particular (and order them in the text by how important I regard them to be). The first question concerns the notion of "medium independence," which is crucial to the mechanistic account. The second question concerns the notion of teleological function that Piccinini introduces and defends.

Before delving into the issues, it is worth pointing out that, in the process of presenting the mechanistic account, Piccinini offers insights on the relationship between functional and mechanistic explanation, on the nature of analogue computation, on the status of connectionism, and he proposes a novel way of understanding teleological function. The book is rich, clear, and truly worth reading.

MEDIUM-INDEPENDENCE

The mechanistic account of computation that Piccinini introduces holds the following: "A physical computation is the manipulation (by a functional mechanism) of a medium-independent vehicle according to a rule. A medium-independent vehicle is a physical variable defined solely in terms of its degrees of freedom (e.g. whether its value is 1 or 0 during a given time interval) as opposed to its specific psychical composition (e.g. whether it's a voltage and what voltage values correspond to 1 or 0 during a given time interval). A rule is a mapping from inputs and/or internal states to internal states and/or output" (10).

Piccinini says that, in this context, a variable is a physical state that can change over time (121, fn. 2). The notion of medium-independence accounts, in the first place, for the fact that computations can be implemented in different physical systems. Medium-independence entails multiple realizability (123). An account of computation that does not allow for multiple realizability—that is, for the same computation being implementable in different physical media—would be a wanting account.

But the notion of medium independence, as stated, is overly liberal. Any physical (or chemical) state that changes over time (it seems) can be *defined in terms of its degrees of freedom* rather than in terms of its specific physical composition. If that is all there is to the notion of medium independence, then it seems that the mechanistic account would, like the mapping account, count too many things (such as stomachs and washing machines) as computing (147 and 275).

Sometimes Piccinini talks of medium-independence in terms of a process acting only on certain physical properties of a state (and not on others). He says: "When we define concrete computations and the vehicles that they manipulate we need not consider all of their specific physical properties. We may consider only the properties that are relevant to the computation, according to the rules that define the computation" (122). But this also seems to be a fairly permissive notion. Despite what Piccinini says (147), the physical states involved in digestion are processed according to some of the physical-chemical properties of the states and not others (they are not processed, for example, according to the color properties they have).

There is clearly further work that the notion of medium independence is required to do. Piccinini says that, although medium-independence entails multiple realizability, the reverse is not the case. This is because medium independence puts structural constraints on the vehicles of computation (123). But what kind of constraints are structural constraints?

We know what the constraints are not. They are not semantic in nature. Medium-independence is invoked to avoid an overly liberal view of computation without appealing to the notion of representation, as the semantic view of computation does. Medium-independent vehicles may, but need not, carry information about something and/or represent it.

We are, however, left wondering what Piccinini means by structural constraints. To illustrate what he means, Piccinini gives the example of digits: "In the case of a medium-independent property, the structural constraint comes from requiring that the medium—any medium that realizes the property—possesses the degrees of freedom that are needed to realize the property. In the case of digits, their defining characteristic is that they are unambiguously distinguishable by the processing mechanism under normal operating conditions. . . . The rules defining digital computations are defined in terms of strings of digits and internal states of the system, which are simply states that the system can distinguish from one another. No further physical properties of a physical medium are relevant to whether they implement digital computations. Thus, digital computations can be implemented by any physical medium with the right degrees of freedom" (123).

As already mentioned, it is unclear how the degrees of freedom requirement would constitute a constraint on a state. On the other hand, the constraints on digits—that they are distinguishable unambiguously by a system throughout a process—do indeed mark a distinctive requirement on computation, but they also restrict computations to only the digital ones. Analogue, quantum, and even (some) connectionist networks would fail to qualify as manipulating medium-independent vehicles because continuous variables and qudits are not unambiguously distinguishable from one another (6 and 124). So the notion of medium-independence needs better unpacking on pain, among other things, of having a mechanistic account that is not superior to competitors in the things it regards as computing.

FUNCTION

Piccinini understands teleological functions as contributions to objective goals of organisms (108). A teleological function in an organism, according to him, is a stable contribution by a trait (or component, activity, property) of organisms belonging to a biological population to an objective goal of those organisms.

This account is supposed to, among other things, avoid the epistemic problems of etiological accounts of function

which identify functions as what historically has served an organism in survival (102). One of the problems with etiological accounts is that "the causal histories that ground functions . . . are often unknown (and in many cases, unknowable), making function attribution difficult or even impossible" (102).

The challenge for Piccinini here is to spell out the notion of a goal, and explain why goals themselves are not historically established. People's goals are, of course, varied and dependent on the present. But in his formulation of function, Piccinini refers to the goals of organisms belonging to a biological population. What are the goals of such organisms and are they knowable? Don't they depend on the evolutionary history of the population? It seems that, absent some further unpacking, the same problems Piccinini raises for teleology apply to his position.

The Mechanistic Account of Physical Computation: Some Clarifications

Gualtiero Piccinini
UNIVERSITY OF MISSOURI–ST. LOUIS

I'm deeply grateful to my commentators for their insightful remarks, which provide me this opportunity to clarify the mechanistic account of physical computation. When I refer to the mechanistic account, I mean the view I defend in my book.[1] The mechanistic account says that computational explanation is a special kind of mechanistic explanation. According to the mechanistic account, physical computation is the manipulation of medium-independent vehicles according to a rule, by a mechanism that is performing one of its teleological functions.

1. HORIZONTAL EXPLANATIONS ARE MECHANISTIC

The mechanistic account maintains that functional analyses are mechanism sketches. This is relevant because computational explanation is traditionally said to be functional analysis. Functional analysis is the explanation of a capacity of a system in terms of organized sub-capacities or sub-functions. Functional analysis is traditionally seen as distinct and autonomous from mechanistic explanation because functional analysis may omit information about components performing the sub-functions.

In order for the mechanistic account to get properly off the ground, the relation between functional analysis and mechanistic explanation must be clarified. I argue that, insofar as functional analysis is less than a complete mechanistic explanation because it omits information about components, it is a mechanism sketch to be completed by adding information about components.[2] (Side note: from the conclusion that functional analysis is a mechanism sketch it doesn't follow that computational explanation is always a mechanism sketch. It may or may not be depending on whether it includes the components.)

Martin Roth (this issue) replies that capacities admit of both horizontal and vertical explanations. Horizontal explanations analyze a capacity in terms of organized sub-capacities; they are functional analyses. Vertical explanations provide implementation details; they are mechanistic explanations. The vertical details can help identify the correct analysis, but they do not add to its explanatory power. Therefore, functional analysis remains explanatorily autonomous from mechanistic explanation.[3]

My point of disagreement is small but crucial. Horizontal explanations involve more than organized sub-capacities; they also involve components possessing those sub-capacities. Capacities plus components amount to mechanistic explanations. Therefore, even horizontal explanations are mechanistic. Since horizontal explanations *are* mechanistic, a fortiori they cannot be autonomous from mechanistic explanations.

Considering an example might help. Roth mentions mousetraps, which are a classic example. How does a horizontal explanation of a mousetrap go? It depends on the kind of mousetrap! There are snap traps, electric traps, glue traps, bucket traps, and more. As their names indicate, different kinds of mousetraps involve different kinds of component (spring-loaded bars, electrical circuits, glue, buckets) possessing different sub-capacities (releasing enough force to kill a mouse, releasing enough electricity to kill a mouse, gluing a mouse down, having walls too deep for a mouse to escape from). There is no such thing as *the* functional analysis of mousetraps. For each type of mousetrap, there is a corresponding horizontal explanation involving specific types of component and their specific sub-capacities.

Sometimes, defenders of functional analysis as distinct and autonomous from mechanistic explanation insist that components can be individuated functionally, by their capacities alone. Even so, my point stands. Any functional individuation puts constraints on the structures that can perform it and, vice versa, there is no structure that can perform all functions. In our example, only certain kinds of structure can function as snaps, electrical circuits, glue, or buckets. That's not to say that all implementational details must be included in a horizontal explanation. I agree with Roth that horizontal explanation abstracts away from, and can be fully explanatory without, lower-level details.[4] Still, horizontal explanation remains mechanistic.

Instead of thinking of implementation as a relation between functional analysis and mechanistic explanation, as Roth seems to do, I think it's more accurate and helpful to think of levels of mechanistic organization implementing one another.[5] Lower levels of mechanistic organization implement higher levels. Higher levels are aspects of lower levels and include the most relevant causes that explain a phenomenon. But this does not lead from mechanistic explanation to functional analysis as a distinct type of explanation; it simply leads from lower to higher levels of mechanistic organization.

One confusing feature of this debate is that, at least in psychology and neuroscience, we tend to focus on *computational* explanation. Computation is *medium independent*, meaning that it abstracts away from most aspects of the physical medium in which the process is realized (except certain degrees of freedom and organizational relations). Even in computational explanation, however, we can't fully explain a capacity, even at the highest level of organization, without specifying the components that perform the relevant sub-capacities. The illusion that we can may be due to the special case in which the same, versatile component performs all the sub-capacities.

In some computational explanations, there is either a central processing unit or a computing human who follows an algorithm by performing all the needed operations. This makes it sound like components don't matter to functional analysis. But we shouldn't let the possibility that one component possesses all the sub-capacities mislead us into thinking that we have stumbled upon a distinct type of explanation. It's just the special case in which one component does all the work. The general type of explanation is still mechanistic, at one level of organization. In this case, there is just one component. Typically, there are many types of component, each of which specializes in one or a few of the sub-capacities that explain the capacity of the system. Either way, horizontal explanations are mechanistic. A fortiori, they are not autonomous from mechanistic explanations.

2.TELEOLOGICAL FUNCTIONS

The mechanistic account relies on the notion of teleological function, which I explicate as a stable contribution to a goal of organisms belonging to a biological population (I will omit "teleological" from now on).[6] In turn, organisms' goals divide into objective and subjective goals. The objective goals of organisms include survival and inclusive fitness. The subjective goals of organisms are those the organisms choose for themselves. This account applies to the functions of both organismic traits and artifacts. Biological traits have the function to contribute to the goals of organisms that possess them. Artifacts have the function to contribute to the goals of organisms that make and use them.

This goal-contribution account has an important advantage over the traditional selectionist account. The selectionist account is that functions are selected effects—effects selected for through a process such as evolution by natural selection.[7] Given the selectionist account, knowing a trait's function requires knowing its selection history. In the case of many biological traits, especially psychological traits, it's very difficult to know their selection history. Yet we often attribute functions correctly without knowing the selection history of the traits that possess them. How do we do it, and what are functions such that we can do it?

The goal-contribution account answers as follows. Functions may or may not be selected for. That's why we don't have to know their selection history in order to attribute them. Instead, what we need to find is the stable contribution that a trait (or artifact) makes to the goals of organisms. That stable contribution is the trait's (or artifact's) function.

Nico Orlandi (this issue) asks, aren't organisms' objective goals dependent on organisms' history, and doesn't this pose the same epistemic challenge for the goal-contribution account that the selectionist account faces? If so, the goal-contribution account has no advantage over the selectionist account.

As I said, the objective goals of organisms include survival and inclusive fitness. These goals are essential to living organisms in the sense that if all organisms cease to pursue them, organisms will leave no descendants and go extinct. This is why survival and inclusive fitness are goals. They are a biological imperative. They must be pursued on pain of extinction.

Since life has a historical origin, we might want to say that survival and inclusive fitness have a historical origin too. In addition, each species evolves its own traits, which allow its members to pursue survival and inclusive fitness in their own species-specific ways. Nevertheless, we do not need to know a species' evolutionary history to discover the specific ways in which, right now, its traits contribute to objective goals such as survival and inclusive fitness. We can discover those contributions by observing and experimenting on current organisms. Because of this, the goal-contribution account of functions retains its advantage over the selectionist account.

3. MEDIUM INDEPENDENCE

The mechanistic account relies on the notion of medium independence, which I explicate as a higher-level property defined solely in terms of degrees of freedom and their organization. For example, digital computations are defined over strings of digits. I define a (physical) digit as a variable that can take a finite number of stable values, which can be reliably distinguished from other values, concatenated into strings, and processed by a physical system in accordance with a rule. This notion of digit is medium independent. It does not specify any more concrete physical properties of digits. Therefore, medium independence is a stronger condition than multiple realizability. Multiple realizability does specify concrete physical properties—for example, *catching mice*. A medium-independent property is multiply realizable but the converse need not hold. For example, being a mousetrap is not a medium-independent property precisely because all mousetraps must handle mice, and mice are a specific type of physical "medium."

In response, Orlandi (this issue) worries that the notion of medium independence is too liberal and does not pose enough constraints on its lower-level realizers. If so, the mechanistic account of computation might count too many systems as computing.

To address this worry, we should realize that the notion of medium independence is not the only constraint on physical computation. For a system to count as computing, it must also be a mechanism with teleological functions and it must manipulate its inputs and internal states in accordance with a rule. The way medium independence comes into play is as follows.

For a mechanism to count as a computing system, its teleological function must be defined in terms of a rule for manipulating medium-independent vehicles. Thus, all the elements of the account work together to constrain the systems involved. The relevant type of medium-independent vehicle defines the structural constraints that must be satisfied by a mechanism in order to count as something that performs computations over such vehicles, provided that it manipulates such vehicles *because* that is its function *and* does so in accordance with a rule.

Okay, but what are the constraints imposed by medium independence? It depends on the type of vehicle. In the case of strings of digits, the constraints include the following. The system must possess components that maintain a finite number of distinguishable stable states. There must be enough components to store all the needed digits. The digits must remain stable for a long enough time that the system can process them successfully. The system must possess components that process such stable states in accordance with the relevant rules. The components must be organized so as to respect the ordering of the strings. Finally, the components must be synchronized so that their states update without disrupting the ordering of the strings. Further constraints are imposed by the functions to be computed and the architecture for computing such functions.

Mutatis mutandis, other types of medium-independent vehicle impose their own constraints on systems that manipulate them. In conclusion, medium independence gives rise to especially abstract levels of mechanistic organization. Nevertheless, computing systems remain pluralities of organized components performing sub-functions in such a way that they produce the capacities of the system they compose. That is, computing systems remain mechanisms.

4. COMPUTING FUNCTIONS

The mechanistic account says that the function of computing mechanisms is to process medium-independent vehicles according to a rule. Frances Egan (this issue) finds that implausible. She points out that computing mechanisms have functions such as parsing, adding, producing speech, etc. She ventures that manipulating medium-independent vehicles is the common *means* by which computing mechanisms perform their functions. That last point is correct, and it's not an accident. Why is it that all computing mechanisms process medium-independent vehicles? Because computing a function is a medium-independent notion. Medium independence is part of the essence of computing; it's a necessary property of all physical computations.

Of course, computational processes also have more specific functions than processing medium-independent vehicles according to a rule. Those are the functions that Egan focuses on: parsing, adding, and so forth. What they all have in common is that they are defined in a medium-independent way.

Here is another way to put the point. Computation is a mathematical notion. The mathematics of computation abstracts away from most physical properties, except whatever degrees of freedom and organizational relations between them are needed to embody an encoding of the function being computed. But that's what medium independence amounts to. Therefore, if a physical system has the function of computing, such a function is defined in a medium-independent way.

But isn't this just a version of the mapping account of physical computation? And doesn't the mechanistic account need the mapping account to tell it which function is being computed? These are the other questions raised by Egan (this issue).

Here is Egan's version of the mapping account:

> [A] physical system S computes a function F just in case:
>
> (i) There is a mapping from (equivalence classes of) *physical states* of S to the arguments and values of F, such that:
>
> (ii) Causal state-transitions between the *physical states* mirror the formal dependency relations between the arguments and values of F. (Egan, this issue, emphasis mine)

Egan argues that the mechanistic account needs to be supplemented by this sort of mapping account in order to identify the function computed by the system. She also argues that this sort of mapping account satisfies two important desiderata. First, it is extensionally adequate—that is, it does not count too many systems as computational. Second, it can account for miscomputation: miscomputation occurs when the system's output fails to map to the relevant value of the function.

Egan is right to this extent: there is a mapping from some of the *computational* states of a physical system to the arguments and values of the function the system computes. She is also right that the system miscomputes when its output fails to map onto the correct value of the function being computed. The problem, which the mapping account does not solve, is identifying the computational states of a physical system. Without that, the relevant mapping cannot be constructed and miscomputation cannot occur.

The mapping account refers to equivalence classes of physical states. Which physical states? Which equivalence classes? Every physical system has lots and lots of microstates, which can be grouped into equivalence classes in many ways. To see why the mapping account is insufficient, consider an arbitrary physical system. Group its microstates into equivalence classes that respect the causal relations between microstates. Label each equivalence class with a symbol from your favorite computational formalism. You have just constructed a mapping from some equivalence classes of physical states to the arguments and values of a function. By respecting the causal relations between microstates, you have prevented lots of spurious computational descriptions and avoided the most pernicious versions of pancomputationalism.

Have you identified a genuine computational system yet? Have you identified the system's computational states? Unfortunately, no. Computational states are equivalence classes of microstates, but most equivalence classes of microstates are not computational states. You were looking for a computational explanation of your system, but all you got is a mere computational *model*. A computational model is not necessarily a computational explanation.[8]

If the system you picked was a paradigmatic noncomputational system—something like a river, a tornado, or a comet—mapping equivalence classes of its microstates to the arguments and values of a function does not give you the computational states of the system, for the system has no computational states at all! It cannot miscompute because it does not compute. Any of its state transitions can be mapped to the arguments and values of a function. All the mapping gives you is a computational model. Just because a system has a computational model, it doesn't follow that the system itself performs computations.

What about cognitive systems? As Egan points out, cognitive scientists begin with specific capacities such as language processing and problem solving, and then they try to find computational explanations for them. Doesn't this lead to genuine computational explanations, which give rise to the possibility of miscomputation? Sure, but it remains to be seen what it takes for a cognitive system, like any other physical system, to be computational. The mapping account says that all it takes is a mapping from equivalence classes of physical states to the arguments and values of a function. That's not enough to identify genuine computational states.

Even in genuine computational systems there are lots and lots of microstates, and equivalence classes thereof, that do not belong to any computational states. For example, the microstates of your computer's battery, fan, and case do not belong to any computational states. What's more, even many microstates of computing components such as logic gates do not belong to any computational states, because they occur at computationally irrelevant times. Therefore, if we are looking for a system's computational states, the last thing we should do is simply to construct a mapping from equivalence classes of microstates to the arguments and values of a function. Before we construct such a mapping, we need to identify the genuine computational states.

The mechanistic account tells you how to identify computational states within physical systems by constructing the relevant mechanistic explanation. First, set aside any system that is not a functional mechanism—any system that lacks (teleological) functions. That rules out rivers, tornadoes, comets, and the like. Second, set aside any system whose functions are not defined in terms of processing vehicles defined in a medium-independent way. That rules out most biological systems and artifacts—stuff like hearts, livers, and vacuum cleaners. Third, set aside any system that does not follow a rule—stuff like random "number" generators. At this point, you've more or less identified the class of physical computing systems. You still have to identify their computational states.

To do that, you have to look at how a computing system performs its primary function—that of processing a medium-independent vehicle in accordance with a rule. If you know the type of vehicle and the rule, you know the function the system computes. But your job is not done. What you need to do is provide a mechanistic explanation of the system's behavior. Where do the inputs enter and the outputs exit? What are the components that store and process the vehicles? How are the vehicles manipulated? More generally, what is the mechanistic organization through which the system performs its computational function? By answering these questions, we can identify genuine computational states and state transitions. Through this process, we can construct genuine computational explanations. We can also identify state transitions between computational states that do not follow the relevant rule, and therefore amount to miscomputation. According to the mechanistic account, this is what cognitive scientists and computer engineers do when they explain how physical systems compute.

In summary, the mapping account is not enough to identify genuine computational states and computational state transitions. It counts too many physical systems as computational and does not really account for miscomputation. Nevertheless, if a physical system is a computing mechanism that is designed correctly, is functioning correctly, is programmed correctly, and is used correctly, there is certainly a mapping from its computational states to the arguments and values of the function it computes. That's the nugget of truth within the mapping account.

Does that mean that the mechanistic account needs to be supplemented by the mapping account in order to identify the function being computed? Not really. The function being computed is a function from input computational states and internal computational states to output computational states. Once the system is explained mechanistically, the function being computed is already included in the computational explanation. There is nothing wrong with pointing out that the computational states map to the arguments and values of a mathematical function, but that does not add any new information to our explanation.

5. QUANTUM MECHANISMS

The mechanistic account relies on the notion of, well, mechanism. A mechanism is a plurality of organized components that, collectively, produce a phenomenon. John Symons (this issue) objects that mechanisms are difficult to square with quantum mechanics, one of our fundamental physical theories. He implies that the mechanistic account is committed to local causal interactions between isolable parts with definite properties. But every physical system, including computing systems, is ultimately made of quantum mechanical systems. In some cases, Symons adds, quantum effects may be relevant to the performance of computing components that are already present within our artifacts. Quantum systems need not have isolable parts, such parts need not have definite properties (because of Heisenberg's uncertainty principle), and the interactions between such parts need not be local (because of entanglement and the wave-particle duality). Therefore,

Symons concludes, insofar as mechanistic explanation requires local causal interactions between isolable parts, it will be unable to explain computation (or anything else, for that matter) once we reach the quantum level.

The literature on mechanisms is vast. Perhaps there are mechanists who are committed to local causal interactions between isolable parts with definite properties. Typical examples of mechanisms in much recent literature are certainly classical (i.e., nonquantum). In my writing, I usually focus on classical (i.e., nonquantum) computation.[9] But my focus on classical computation does not imply that quantum systems cannot be mechanistic. The notion of mechanism is flexible and open ended enough to encompass quantum systems. After all, it's quantum *mechanics*! Several issues should be distinguished.

A first issue is that although everything is made out of quantum mechanical parts, typical macroscopic objects behave classically. Why is that? A plausible explanation is decoherence.[10] Decoherence occurs when a complex quantum mechanical system interacts with enough portions of its environment to become entangled with them. Enough entanglements with enough things make the superposition between quantum states undetectable and suppresses exotic quantum effects such as quantum mechanical interference. That's probably why most of the objects we observe, including ordinary computers, behave classically. As long as a system and its components behave classically, quantum mechanics is not especially relevant and poses no special problem.

Eventually, though, if we keep explaining a phenomenon mechanistically and descend to more and more microscopic levels, we will reach a level at which quantum effects become relevant. At this stage, we can appeal to what Kuhlmann and Glennan call *nonclassical mechanistic explanation*.[11] For present purposes, nonclassical mechanistic explanation is explanation of a phenomenon in terms of components that exhibit nonclassical features, such as parts that are not isolable from one another and may even lack definite locations, lack definite properties (in the sense that they satisfy Heisenber's uncertainty principle), etc. As to nonlocal effects due to entanglement, they often play no role in a quantum system's dynamics. When entanglement affects dynamics, Kuhlman and Glennan argue, we finally meet the limits of the mechanistic program. Nonlocal effects due to quantum entanglement lack a mechanistic explanation. Another place where mechanistic explanation stops is where we find truly basic components—elementary particles. Since they lack parts, their behavior has no mechanistic explanation. This is not a flaw: mechanistic explanation, like explanation simpliciter, has to stop somewhere. Mechanistic explanation stops when there are no smaller parts whose sub-capacities explain the capacities of the whole.

Nevertheless, nonlocal effects and any other basic physical actions can partake in mechanistic explanation. An especially relevant case is quantum computation. Quantum computation is like classical digital computation except that, instead of operating on digits (typically, bits), it operates on qudits (typically, qubits). Roughly, qudits are d-dimensional quantum systems which can be put into superpositions of computational basis states and also be entangled with one another. In some cases, exploiting these quantum mechanical features can lead to greater efficiency in the computation.

Armond Duwell has looked at an especially exotic type of quantum computing systems: Measurement Based Quantum Computers (MBQCs).[12] In addition to superposition, MBQCs exploit entanglements between qubits. Unlike ordinary classical *and* quantum computing systems, in which digits or qubits flow through a circuit, MBQCs operate by performing a series of measurements on an array of entangled qubits. In spite of the unusual and highly nonclassical structure of MBQCs, Duwell shows how to apply the mechanistic account to them.

Duwell points out that, even though qubits do not flow through MBQCs the way they flow through more ordinary computing systems, the qubits in a MBQC array are correlated in such ways, due to their entanglement, that a flow of qubits is not necessary to complete a computation. Instead, the whole array of qubits is the computational vehicle and, given the entanglement of the qubits, measurements affect not only the measured qubits but also all the qubits entangled with them. What's more, it is the *function* of the measurement apparatus to affect the qubits in this way, and it is a *function* of the entanglement of the qubits to be affected in this way so as to produce the correct output. Given that other components of the system have other functions that contribute to the computation, the whole system is a functional mechanism whose function is processing a medium-independent vehicle (qubit array) in accordance with a rule. Thus, the mechanistic account applies to MBQCs.

In conclusion, quantum mechanics requires adjustments to the mechanistic account of physical computation but is compatible with it.

6. CONCLUSION

Thanks to the thoughtful commentaries by Roth, Orlandi, Egan, and Symons, I did my best to clarify why I believe functional analyses are mechanism sketches, teleological functions are regular contributions to the goals of organisms, computation is medium independent, the function of computing mechanisms is to process medium-independent vehicles according to a rule, and quantum systems can be nonclassical mechanisms. There is room for more detailed philosophical work on these topics. I greatly appreciate this opportunity to get started. I hope others will contribute as well.

ACKNOWLEDGMENTS

Thanks to Marcello Guarini for organizing the APA session from which this exchange derives, and to Peter Boltuc for inviting us to publish our contributions in this newsletter. Thanks to Armond Duwell for help with quantum computation.

NOTES

1. Piccinini, *Physical Computation: A Mechanistic Account*; for related accounts, see Kaplan, "Explanation and Description in Computational Neuroscience"; Fresco, *Physical Computation and Cognitive Science*; Milkowski, *Explaining the Computational*

Mind; Coelho Mollo, "Functional Individuation, Mechanistic Implementation: The Proper Way of Seeing the Mechanistic View of Concrete Computation."

2. Piccinini, *Physical Computation*, Chap. 5; Piccinini and Craver, "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches."
3. Cf. Roth and Cummins, "Neuroscience, Psychology, Reduction, and Functional Analysis."
4. Boone and Piccinini, "Mechanistic Abstraction."
5. Boone and Piccinini, "The Cognitive Neuroscience Revolution."
6. Maley and Piccinini, "A Unified Mechanistic Account of Teleological Functions for Psychology and Neuroscience."
7. Garson,*What Biological Functions Are and Why They Matter.*
8. Piccinini,*Physical Computation*, Chap. 4.
9. The notion of classical computation as opposed to quantum should not be confused with the notion of classical computation in the sense of digital computation operating on language-like vehicles (Fodor and Pylyshyn, "Connectionism and Cognitive Architecture"). Here we are discussing classical versus quantum mechanical physical systems and their processes.
10. Kuhlmann and Glennan, "On the Relation Between Quantum Mechanical and Neo-Mechanistic Ontologies and Explanatory Strategies."
11. Ibid., Section 5.
12. Duwell, "Exploring the Frontiers of Computation: Measurement Based Quantum Computers and the Mechanistic View of Computation."

REFERENCES

Boone, W., and G. Piccinini. "Mechanistic Abstraction." *Philosophy of Science* 83, no. 5 (2016a): 686–97.

Boone, W., and G. Piccinini. "The Cognitive Neuroscience Revolution." *Synthese* 193, no. 5 (2016b): 1509–34.

Coelho Mollo, D. "Functional Individuation, Mechanistic Implementation: The Proper Way of Seeing the Mechanistic View of Concrete Computation." *Synthese* 195, no. 8 (2018): 3477–97.

Duwell, A. "Exploring the Frontiers of Computation: Measurement Based Quantum Computers and the Mechanistic View of Computation." In *Turing 100: Philosophical Explorations of the Legacy of Alan Turing*, edited by A. Bokulich and J. Floyd. Boston Studies in the Philosophy and History of Science, vol. 324, pp. 219–32. Springer, 2017.

Fodor, J. A., and Z. W. Pylyshyn. "Connectionism and Cognitive Architecture." *Cognition* 28 (1988).

Fresco, N. *Physical Computation and Cognitive Science*. New York: Springer, 2013.

Garson, J. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press, 2019.

Kaplan, D. M. "Explanation and Description in Computational Neuroscience." *Synthese* 183, no. 3 (2011): 339–73.

Kuhlmann, M., and Glennan, S. "On the Relation Between Quantum Mechanical and Neo-Mechanistic Ontologies and Explanatory Strategies." *European Journal for Philosophy of Science* 4, no. 3 (2014): 337–59.

Maley, C. J., and G. Piccinini. "A Unified Mechanistic Account of Teleological Functions for Psychology and Neuroscience." In *Explanation and Integration in Mind and Brain Science*, edited by D. Kaplan, 236–56. Oxford: Oxford University Press, 2017.

Milkowski, M. *Explaining the Computational Mind*. Cambridge, MA: MIT Press, 2013.

Piccinini, G. *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press, 2015.

Piccinini, G., and C. Craver. "Integrating Psychology and Neurscience: Functional Analyses as Mechanism Sketches." *Synthese* 183, no. 3 (2011): 283–311.

Roth, M., and R. Cummins. "Neuroscience, Psychology, Reduction, and Functional Analysis." In *Explanation and Integration in Mind and Brain Science*, edited by D. Kaplan. Oxford: Oxford University Press, 2017.

Philosophical Insights from Computational Studies: Why Should Computational Thinking Matter to Philosophers?

Gary Mar

STONY BROOK UNIVERSITY

(With Edward Zalta and Aydin Mohseni)

In April the Committee on Philosophy and Computers sponsored two panels at the APA Pacific meeting in Vancouver, Canada. The panel on "Data Ethics" was chaired by Joshua August Skorburg, who also delivered a paper along with papers by Shannon Vallor and Colin Koopman. The second panel, "Philosophical Insights from Computational Studies: Why Should Computational Thinking Matter to Philosophers?," was chaired by Gary Mar, who delivered a paper along with papers by Aydin Mohseni and by Edward Zalta. What follows are two abstracts and one summary of the latter session.

Despite being scheduled during the next to the last session on the last day of the conference, the panel on "Philosophical Insights from Computational Studies" was well attended and provoked a spirited debate. This is perhaps evidence that the APA Committee on Philosophy and Computers has progressed well beyond raising pedagogical questions about the use of technology in the classroom and is now, as it is soon to be disbanded, addressing fundamental questions about how computationalism is profoundly transforming the nature philosophical research and the kinds of questions that can now be addressed.

1. ON THE EMERGENCE OF MINORITY DISADVANTAGE: TESTING THE *RED KING HYPOTHESIS* (REPORT WITH ABSTRACT).

Aydin Mohseni, a graduate student in the Department of Logic and Philosophy of Science at the University of California, Irvine after having completed an MA in Logic, Computation, and Methodology at Carnegie Mellon University, presenting joint research conducted with Cailin O'Connor (University of California, Irvine) and Hannah Rubin (Notre Dame University).

The cultural Red King effect was first described by political philosopher Justin Bruner, who uses evolutionary game theoretic methods to show how minority groups can be disadvantaged in the emergence of bargaining conventions solely by dint of their group size.[1] As he shows, in groups with completely symmetric preferences, abilities, and resources, minority status alone can increase the likelihood that individuals end up with fewer economic resources. The driver behind this effect is a learning asymmetry between minority and majority groups. While minority members commonly meet their out-group, the reverse is not true. As a result, members of a minority will more quickly learn to interact with their out-group. In situations where this learning is about bargaining interactions, this often proves

disadvantageous. Low, accommodating demands tend to be safer in bargaining interactions, meaning that swift learners should adopt these demands. Once this is done, members of the majority group can take advantage of this accommodation.

Subsequent work has shown that this effect arises robustly in cultural evolutionary models.[2] Given the simplicity of these models, though, a further question arises: Can the cultural Red King really occur in human groups? If so, there are important consequences for social and political philosophy. Future work will be directed at better understanding the empirical conditions under which cultural red king-type effects may obtain and contribute to the emergence of inequitable conventions.

The populations are playing a Nash demand game with two strategies: aggressive demand, and passive demand. The x-axis reflects the proportion of individuals in the first population who are playing each strategy while the y-axis reflects the proportion in the other population. There are two attracting states corresponding to the possible outcomes of evolution: the state where the first population adopts the aggressive strategy and the second population the passive strategy, and the dual state where the strategies are flipped. When the two populations' rates of evolution are equal, the basins of attraction for each attractor are symmetric (shown in Figure 1). However, when the rates of evolution are unequal, the basin of attraction for the attractor where the slower evolving population plays the aggressive strategy grows larger (show in the right figure). This is the red king effect: the slower runner wins the race. The cultural red king effect then consists in observing that a functionally equivalent effect can be produced via population size differences conducing to differences in rates of interaction and so to learning rates.

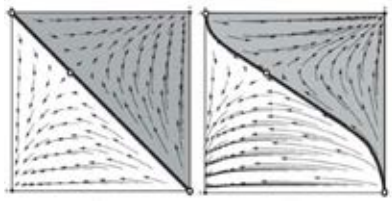


Figure 1. The phase portrait captures the dynamics of two evolving populations as they interact.

2. METAPHYSICAL INSIGHTS FROM COMPUTATIONAL STUDIES (EXTENDED ABSTRACT)

Edward Zalta, Senior Research Scholar at Stanford's Center for the Study of Language and Information (CSLI) and the winner of the 2016 Jon Barwise Prize.

The foundations of object theory have evolved in a number of interesting new ways since its implementation in Isabelle/HOL by Christoph Benzmueller and Daniel Kirchner. Not only has a potential "back-door paradox" (i.e., a loophole permits the reintroduction of a known paradox) been identified and avoided, but the language of object theory itself has been improved by eliminating the syntactic category of "propositional formulas" (i.e., formulas containing encoding subformulas). λ-expressions $[\lambda x_1 \ldots x_n\ \phi]$ whose matrix ϕ has encoding subformulas are now well-formed, though they are guaranteed to denote relations only if ϕ is encoding- and description-free. Moreover, the language of object theory has been generalized to allow n-ary encoding formulas. These changes to the language have led to changes in the definitions, axioms, and theorems of the theory, primarily by extending the free logic for definite descriptions to λ-expressions. Two new axioms have been added: one ensures that any relation necessarily equivalent to an existing relation also exists, and a second ensures that a formula denotes a proposition precisely when its nominalization denotes a proposition. Finally, the axiom for asserting that there are contingently nonconcrete objects has been refined.

These changes have metaphysical implications:

1. Not only can identity for objects and relations be defined in terms of predication, but existence can be defined for objects and relations in terms of predication as well. (By contrast, free logics either reduce existence to a primitive notion of identity or take existence as a primitive.)

2. The new axioms also allow us to prove that *every* formula denotes a proposition (though not every open formula can be transformed into a denoting λ-expression). This, in turn, extends the fundamental theorem of world theory, which previously applied only to formulas without encoding subformulas: the fundamental theorem now governs every formula whatsoever: necessarily ϕ if and only if ϕ is true in every possible world.

3. Finally, with a refined axiom that asserts the *possible* existence of a nonconcrete but actually concrete object, the theory still guarantees that at least two possible worlds and four propositions exist, but also that there are at least 16 properties.

3. GÖDEL, TURING, AND TIME: A COMPUTATIONAL PHILOSOPHY OF MATHEMATICS (SUMMARY)

The philosophical views of Kurt Gödel and Alan Turing are often presented as posing an irreconcilable dichotomy: Gödel's *platonism* asserts that the human mind is creative surpassing the capacity of any single Turing machine whereas Turing's *computationalism* leads him to the mechanistic view that the human brain is essentially a Turing machine. Gödel and Turing, however, are more philosophically honest and skeptical than their subsequent followers who have posed the dichotomy as a stereotypical choice between the *creative mind* versus *mechanistic computability* in the philosophy of mathematics. It is argued that whereas Gödel and Turing's views on the philosophy of mind might be incompatible, their philosophies of *mathematics* are not irreconcilably inconsistent but, in fact, are complementary.

This thesis faces three immediate objections. First, platonism is about abstract, universal, and *timeless objects* whereas computations are *processes* that take place *in time*. In his conversations with Hao Wang, Gödel remarked:

> The real argument for objectivism is the following. We know many general propositions about natural numbers to be true (2 plus 2 is 4, there are infinitely many prime numbers, etc.) and, for example, we believe that Goldbach's conjuncture makes sense, must be either true or false without there being any room for arbitrary convention. Hence, there must be objective facts about natural numbers. *But these objective facts must refer to objects that are different from physical objects because, among other things they are unchangeable in time*.[3]

Secondly, in his Josiah Willard Gibbs Lecture "Some Basic Theorems on the Foundations of Mathematics and their Implications" to the meeting of the American Mathematical Society at Brown University in 1951, Gödel proposed the following disjunctive conclusion as "inevitable":

> *Either mathematics is incompleteable in this sense, that its evidence axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there are absolute unsolvable Diophantine problems. . . ."*[4]

Thirdly, Turing's PhD thesis *"Systems of Logic Based on Ordinals"* under the direction of Alonzo Church attempts to overcome the essential incompleteness of Gödel's theorem, the basis for Gödel's disjunction. Turing notes that "in pre-Gödel times it was thought by some that it would probably be possible to . . . [replace] intuitive judgements . . . by a finite number of . . . rules [of formal logic]. The necessity of intuition would then be entirely eliminated."[5] Gödel showed for that any formal system S powerful enough to represent arithmetic, there is a statement G which is true but which the system is unable to prove. Turing's thesis considers the possibility of adding G as an additional axiom to the system and then iterating the process to infinity, creating a system with an infinite set of axioms. Turing explains his motivation:

> In our discussions, however, we have gone to the opposite extreme and eliminated not intuition but ingenuity, and this in spite of the fact that our aim has been in much the same direction. We have been trying to see how far it is possible to eliminate intuition, and leave only ingenuity. We do not mind how much ingenuity is required, and therefore assume it to be available in unlimited supply. In our metamathematical discussions we actually express this assumption rather differently. We are always able to obtain from the rules of a formal a method for enumerating the propositions proved by its means. We then imagine that all proofs take the form of a search through this enumeration for the theorem for which a proof is desired. In this way ingenuity is replaced by patience."[6]

On the contrary, Gödel's and Turing's philosophies of mathematics are not irreconcilably *inconsistent* but *complementary*. This thesis can be proved in five ways. Gödel's endorsement of platonism in print begins with his "Russell's Mathematical Logic" and is reaffirmed in "What Is Cantor's Continuum Problem?" The 1964 supplement for the latter contains Gödel's most full-fledged, frequently quoted, espousal of platonism. Alan Turing's *computational* point of view is contained in his two most famous articles "On Uncomputable Numbers with an Application to the *Entscheidungsproblem*" and "Computing Machinery and Intelligence," that sets forth the now famous Turing Test for answering the question "Can computers think?"

First, Gödel and Turing were "better philosophers" than their followers insofar as their published writings express views that are more nuanced and sceptical than the presentations of their views by ardent and partisan (especially, as regards their views on the philosophy of mind) followers. In his invited lecture "The present situation in the foundations of mathematics" to a joint meeting of the Mathematical Association of America and the American Mathematical Association in Cambridge, Massachusetts, Gödel admitted:

> The result of the preceding discussion is that our axioms, if interpreted as meaningful statements, necessarily presuppose a kind of Platonism, which cannot satisfy any critical mind and which does not even produce the conviction that they are consistent.[7]

Turing's Lecture for *London Mathematical Society* emphasizes the growing importance of human interest in the philosophy of mathematics:

> As regards mathematical philosophy, since the machines will be doing and more mathematics themselves, the centre of gravity of the human interest will be driven further and further into philosophical questions of what can in principle be done etc.[8]

In his introductory remarks to Turing's "Computing Machinery and Intelligence" in the Cooper and Leeuwen Centenary anthology *Alan Turing: His Work and Impact*, Gregory Chaitin in "Mechanical Intelligence versus Uncomputable Creativity" notes: "It is a delightful paradox that Turing argues that we are machines while all the while emphasizing the importance of what machines cannot do. Like a good philosopher, he cannot help seeing the good arguments on both sides. He thus provides ammunition to both parties."[9]

Secondly, a computationalist pedadogy can provide an account of the learnability of mathematics, which poses a central problem for platonism. The confirmed platonist G. H. Hardy described

> the function of a mathematician [as] . . . simply . . . observ[ing] the facts about his own intricate system of reality, that astonishingly beautiful complex of logical relations which forms the subject-matter of his science, as if he were an explorer looking

> at a distant range of mountains, and to record the results of his observations in a series of maps, each of which is a branch of pure mathematics. . . .[10]

In support of his view, Hardy liked to tell a famous anecdote:

> He [Ramanujan] could remember the idiosyncrasies of numbers in an almost uncanny way. It was Littlewood who said that every positive integer was one of Ramanujan's personal friends. I remember once going to see him when he was ill at Putney. I had ridden in taxi cab number 1729 and remarked that the number seemed to me rather a dull one, and that I hoped it was not an unfavorable omen. "No," he replied, "it is a very interesting number; it is the smallest number expressible as the sum of two [positive] cubes in two different ways."[11]

A computational account of Ramanujan's insights is a more plausible and pedagogically sound alternative to Hardy and Littlewood's mystical musings. If you compute a table of cubes for the first dozen integers perhaps you can discover Ramanujan's Taxicab number for yourself:

1	1	**5**	125	**9**	729
2	8	**6**	216	**10**	1000
3	27	**7**	343	**11**	1331
4	64	**8**	512	**12**	1728

Thirdly, mathematical platonism traces its lineage back to the *Meno*, which raises well-known paradoxes about the learnability of mathematics. Philosophically, platonism is about abstract, universal, timeless *objects* whereas computations are *calculations*, processes that take place in time. This dichotomy leads to two fundamentally different approaches to the mathematical enterprise. Philip Davis and Reuben Hersh in *The Mathematical Experience* draw a distinction between *dialectical* {or what I shall call *deductivist*] and *algorithmic* mathematics.[12] According to the former conception, mathematics is a rigorous logical science in which propositions about platonic objects are either true or false. According to the latter conception, mathematics is a tool for solving problems by constructing algorithms. With the increasing use of computers, there has been a shift in mathematics from the former conception back to a more constructive or algorithmic point of view.

These two approaches can be illustrated from the first crisis in the foundations of mathematics—the Pythagorean discovery of that the *diagonal* of the unit square (the answer to Socrates's question and the basis of the geometrical theorem in the *Meno*) is *alogon* or irrational. According to *a priori* deductive mathematics, the mathematical enterprise is to discover proofs about eternal mathematical truths and objects (e.g., that the diagonal of the unit square is irrational).

Theorem: $\sqrt{2}$ is irrational.

Proof: Assume that $\sqrt{2} = a/b$, where a and b are reduced to lowest terms.

1. Hence, $2b^2 = a^2$.
2. $2b^2 = a^2$ implies that a is even so $a = 2c$.
3. Hence, $2b^2 = (2c)^2$ and so $b^2 = 2c^2$, which means that b is even.
4. Contradiction. Both a and b being even contradicts a/b was reduced to lowest terms.

The classical proof of this fact—the proof of a negative existential—is what the Cambridge mathematician G. H. Hardy called a "theorem of the first class." It is interesting to note that a contradiction appears already in step 2, but the continuation of the proof to show the recursive nature of the contradiction is mathematically more elegant.

Figure 2. The Babylonian clay tablet YBC 7289 (c. 1800–1600 B.C.) gives an approximation of the square root of 2 in four sexagesimal figures, 1 24 51 10 = 1.414213, which is accurate to six decimal digits.

According an empirical and computational point of view, the mathematical enterprise is about producing calculations to answer mathematical questions (e.g., how can one calculate with increasing accuracy the length of the diagonal of the unit square?) The approach of *algorithmic*, in contrast to *deductivist*, mathematics is not to prove the *non-existence* of a rational representation for $\sqrt{2}$, but to produce an algorithm for converging on is actual value of $\sqrt{2}$. Such an approach is more practically useful than the purely theoretical platonic approach if, for example, you want to build a bridge.

One such algorithm is known as *Newton's method*. If $x^2 = 2$, then we may divide both sides of the equation by 2 to obtain

$$x = 2/x \ .$$

Let's say our estimate for x is slightly incorrect, say *underestimated*, then $2/x$ will be *overestimated*, and vice versa. Therefore, a better estimate than either x or $x/2$ would be their *average*.

The guiding idea behind Newton's method is captured in the following dynamical system:

$$x_{n+1} = 1/2(x_n + 2/x_n)$$

Newton's method converges quickly. From an initial estimate of 1, for example, we obtain in just four iterations an estimated value that is accurate to 9 digits. This method was known to the ancient Babylonians.

Reflecting for a moment on the *movement* within the classical proof of the irrationality of √2, we may wish to countenance not just the *negative existential* conclusion but the logical *dynamics* of the proof. Here there is a geometric connection to the form of proof by induction, loved by Fermat, known as *proof by infinite descent*. As can be easily seen from their symbolic representations, Fermat's Proof (or Disproof) by Infinite Descent is logically equivalent to Strong Mathematical Induction for natural numbers:

Strong Induction: if F holds of *x* whenever it holds for all numbers less than *x*, then F holds for *all* numbers:

$$\forall x[N(x) \wedge \forall \mathbf{y[N(y) \wedge y < x \rightarrow F(y)]} \rightarrow \mathbf{F(x)}] \rightarrow \forall x[N(x) \rightarrow F(x)]$$

Fermat's Proof by Infinite Descent: F if holding of number implies it holds for an even smaller number, then no number has F:

$$\forall x[N(x) \wedge \mathbf{F(x)} \rightarrow \mathbf{\exists y[N(y) \wedge y < x \wedge F(y)]} \rightarrow \forall x[N(x) \rightarrow \mathbf{\sim} F(x)]$$

The classical proof shows that if √2 *were* rational, no "smallest" representation as a fraction could exist (i.e., a fraction reduced to lowest terms). Any attempt to find a "smallest" representation *a*/*b* would imply the existence of a smaller one, which is impossible given the nature of natural numbers. Fermat's proof by means of the impossibility of method of infinite descent connects a *computational* way of thinking of irrationality with *fractal* geometry—a method that deploys an *infinite-regress* in much the same way as Zeno's paradoxes.

Proof by Infinite Descent. The following diagram was used by Tennenbaum (1950s) to prove that that is irrational by an argument by "infinite descent."

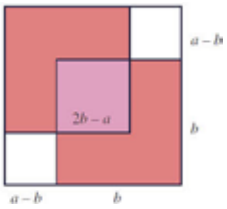


Figure 3. A fractal proof of the irrationality of √2 is a geometric counterpart to Fermat's inductive proofs by the impossibility of infinite descent.

If the diagonal of the unit square is rational, then we have $\sqrt{2} = (a/b)$, where *a* and *b* are the smallest such units with no common factors. This implies that

$$2b^2 = a^2.$$

Geometrically, this means that the area of the square with side *a* is equal to two squares with side *b*. Place the two squares with side *b* inside the square with side *a*. Since the areas of the two *b* squares are equal to the big *a* square, by assumption, the light pink square in the center created by the overlapping *b* squares must be equal in area to the two smaller white squares, which are the remainders uncovered by the overlapping *b* squares. Notice we have that the pink square must be equal to the areas of the two white squares, which provides a smaller solution to our original problem. Contradiction by infinite regress.

This geometric proof by infinite regression foreshadows its fractal nature and relates dynamical systems which can be used to model the paradoxes to Zeno's paradoxes. The seeds of this approach that combines platonism with computationalism can be illustrated by transforming the diagram in the *Meno* into a dynamic construction. Take the unit square from the Meno and connect the midpoints of the square to construct an interior diamond. The area of this diamond is ½ the original square. Then connect the midpoints of the diamond to construct an interior square, which is ½ the area of diamond or ¼ the original square. Keep repeating this construction. What is the sum of the areas of all the nested diamonds and squares? This transformation produces a fractal that geometrically represents the infinite geometric series characteristic of Zeno's dichotomy paradoxes.

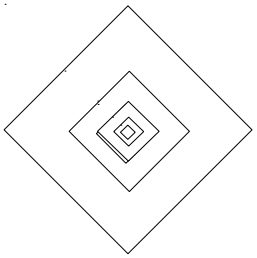


Figure 4. The diagram from the Plato's *Meno* can dynamically be transformed into a fractal representing geometric infinite series of Zeno's dichotomy.

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots .$$

This construction is universal. If you begin with any quadrilateral with area equal to 1, one obtains a nested series of parallelograms whose infinite sum areas is given by Zeno's dichotomous geometric series.

Fourthly, Gödel's platonism (i.e., the objectifying of computations) is *required* for Turing's meta-logical proof of the unsolvability of the halting problem, whereas a kind of platonism (i.e., treating an algorithm as a universal) is *presupposed* in the very conception of the research program of AI. Indeed, the dialectic of informal intuition and formalized algorithmic proof is characteristic of theoretical progress in logic. We can state the mutual relationship between platonism and computationalism in a Kantian manner:

Platonism without computationalism
is epistemologically empty;

Computationalism without platonism
is theoretically blind.

Russell's paradox about the barber who shaves all and only those who don't shave themselves is often treated timelessly. When the problem is stated dynamically about a resident barber in Hilbert's Aleph-Nought Hotel who shaves all and only those residents who have terminating algorithmic schedules, we can obtain a new proof of Turing's Unsolvability of the Halting Problem.

A central thesis of AI can be stated as the analogy that minds are to brains, what software is to hardware. The claim that the same software or algorithm can be instantiated in different brains or in computers presupposes that algorithms not be conceived merely mechanical processes that occur in time but as abstract objects that can be instantiated at different times in different substrata.

Fifthly, combining Gödelian platonism with Turing's computationalism results in a dynamic approach to the semantic paradoxes. Exploring the paradoxes computationally reveals fractal images in the semantics of paradox.[13] Aristotle's law of non-contradiction states that a given proposition cannot be true and false in the same respect *at the same time*. This approach allows for both Gödelian limitative theorems and the discovery of a menagerie of infinitely complex and chaotic paradoxes:

> Paradox is not illogicality, but it has been a trap for logicians: the semantic paradoxes look just a little simpler and more predictable than they actually are. Even in some of the most recent and logically sophisticated work on cyclical regularity in the semantic paradoxes, their deeper and more complex semantic patterns have remained hidden. Our attempt, rather than search for semantic stability or simple patterns within the paradoxes, has been to offer glimpses of the infinitely complex, chaotic, and fractal patterns of semantic instability that have gone virtually unexplored.[14]

The logic and philosophy of time is currently experiencing a renaissance across the disciplines because of the widespread use of computers and algorithms to reframe research questions in philosophy. The semantic paradoxes such as the Paradox of the Liar or the Epimenides Paradox seemed to be an entirely different kind of paradox from Zeno's paradoxes of motion. For example, the former *semantic* paradoxes and have values that are *discrete* and *bivalent*, where the latter paradoxes of motion presuppose an infinity of values—either a *countable* infinity such as that required by the assumption of *infinite divisibility* or the infinity of the *continuum* required by *continuous* motion. Solutions to the semantic paradoxes belongs to the *philosophy of logic*, whereas solutions to Zeno's paradoxes belong to the *philosophy of space and time*. Treating paradoxes *computationally* and *dynamically* provides a *unification* between the semantic paradoxes of the liar in the philosophy of logic and Zeno's paradoxes of motion in the philosophy of space and time.

Replies to Objection #1. Classically, platonism is about abstract, universal, and *timeless objects* whereas computations are *processes* that take place *in time*. However, it has been argued the metatheory of computability requires treating algorithms platonically and timelessly, whereas resolving the epistemological paradoxes about learnability for platonism can benefit from reflection upon processes like calculations that take place in time.

Replies to Objection #2. With characteristic caution, Gödel sometimes prefaced his disjunction with a concession to the possibility of a mechanistic view of mind (*italics mine*):

> The human mind is incapable of formulating (or mechanizing) all its mathematical intuitions. I.e.: If it has succeeded in formulating some of them, this very fact yields new intuitive knowledge, e.g. the consistency of this formalism. This fact may be called the "incompletability" of mathematics.
>
> *On the other hand, on the basis of what has been proved so far, it remains possible that there may exist (and even be empirically discoverable) a theorem-proving machine which in fact is equivalent to mathematical intuition, but cannot be proved to be so, nor even* proved *to yield only correct theorems of finitary number theory.*
>
> The second result is the following disjunction: *Either the human mind surpasses all machines (to be more precise: it can decide more number-theoretic questions than any machine) or else there exist number theoretic questions undecidable for the human mind.*

Replies to Objection #3. It should be noted that both Gödel and Turing would have agreed that no *single* machine is sufficient to simulate the mathematician's mind because of Gödel *incompletability*:

> There would be no question of triumphing simultaneously over all machines. In short, then, there might be men cleverer than any given machine, but then there might be other machines cleverer again, and so on.[15]

Although Turing's attempt in his PhD to overcome Gödel incompleteness still appeals to ingenuity, which is not formalized, in "Intelligent Machinery," Turing envisions an iterative process of mechanization in which a machine takes the initiative to increasingly incorporate the "residue" of human intuition not previously captured:

> If the untrained infant's mind is to become an intelligent one, it must acquire both discipline and initiative. So far, we have been considered only discipline [via the universal machine]. . . . But discipline is certainly not enough in itself to produce intelligence. That which is required in addition we call initiative. This statement will have to serve as a definition. Our task is to discover the nature of this residue as it occurs in man and try to copy it in machines.[16]

Here one is reminded of the Gödel program of the search for increasingly more general Axioms of Infinity to settle questions of set theory.

Gödel rejected mechanistic reductionism for the mind and claimed that Turing's analysis committed a *philosophical* error:

> *A philosophical error in Turing's work*. Turing in his 1937 . . . gives an argument which is supposed to show that the mental procedures cannot go beyond mechanical procedures. However, this argument is inconclusive. What Turing disregards completely is that fact that *mind, in its use, is not static, but constantly developing*, i.e., using them, and that more and more abstract terms enter in the sphere of our understanding. There may exist systematic methods of actualizing this development, which could form part of the procedure. Therefore, although at each stage the

> number and precision of the abstract terms at our disposal may be *finite*, both (and, therefore, also Turing's number of *distinguishable states of mind*) may converge toward infinity in the course of the application of this procedure. Note that something like this indeed seems to happen in the process of forming stronger and stronger axioms of infinity in set theory. This process, however, today is far from being sufficiently understood to form a well-defined procedure. It must be admitted that the construction of a well-defined procedure which could actually be carried out (and would yield a non-recursive number-theoretic function) would require a substantial advance in our understanding of the basic concepts of mathematics.[17]

The scholarly consensus seems to be that Gödel was hasty in attributing the static view of the mind to Turing," where the parentheses indicate that Gödel may have been wrestling with this question as a challenge, not merely to Turing, but to his own views on the matter).[18] Comparing Gödel's notes from 1972 and 1974, Sieg notes some substantive differences and carefully concludes:

> I don't fully understand these enigmatic observations, but three points can be made. First mathematical experience has to be invoked when asking the right questions; second, aspects of that experience may be codified in a mechanical procedure and serve as the basis for asking the right questions; third, the answers may involve abstract terms that are introduced by the nonmechanical mental procedure. We should not dismiss or disregard Gödel's methodological remarks that "asking the right questions on the basis of a mechanical procedure" may be part of a systematic method to push forward the development of the mind. Even this every limited understanding allows us to see that Gödel's reflections overlap with Turing's proposal for investigating matters in a broadly empirical and directly computational manner.[19]

Combining Gödel's platonistic views with Turing's algorithmic methods in a dialectical and recursive manner provides a means for constructing a hybrid computational philosophy of mathematics.

NOTES

1. Bruner, "Minority (dis) Advantage in Population Games."
2. O'Connor, "The Cultural Red King Effect"; O'Connor and Bruner, "Dynamics and Diversity in Epistemic Communities."
3. Wang,*A Logical Journey: From Gödel to Philosophy*, 7.1.2., 211; italics mine.
4. Gödel, "Some Basic Theorems on the Foundations of Mathematics and Their Implications," in GCW-III, 310.
5. Turing, "Systems of Logic Based on Ordinals," 82.
6. Ibid.
7. Gödel [*1933o], in GCW-III, 50.
8. Turing's 1947 lecture for London Mathematical Society.
9. Gregory Chaitin in "Mechanical Intelligence versus Uncomputable Creativity," 551.
10. Hardy, "The Theory of Numbers," *Nature* 110 (Sept. 16, 1922): 381.
11. Ibid., lvii-lviii.
12. Philip Davis and Reuben Hersh, *The Mathematical Experience*, 199.
13. See Ian Stewart's "A Partly True Story" in *Scientific American*, a popular exposition of research in *Philosophical Computer* [1998]).
14. Grim, Mar, and St. Denis, *The Philosophical Computer*, 87. The first sentence is an allusion to G. K. Chesterton, the prince of the paradoxical aphorism.
15. Turing, "Computing Machinery and Intelligence," 445.
16. Turing, "Intelligent Machinery," 21.
17. GCW-II, p. 306.
18. E.g., see Copeland and Shagrir, "Turing versus Gödel on Computability and the Mind"; and Seig, "Gödel's Philosophical Challenge (to Turing)."
19. Seig, "Gödel's Philosophical Challenge (to Turing)," 193.

REFERENCES

Bruner, J. P. "Minority (dis) Advantage in Population Games." *Synthese* (2017): 1–15.

Copeland, B. Jack, ed. *The Essential Turing: The Ideas that Gave Birth to the Computer Age*. Oxford: Clarendon Press, 2004.

Copeland, B. Jack, and Oron Shagrir. "Turing versus Gödel on Computability and the Mind." In *Computability: Turing, Gödel, Church, and Beyond*, edited by B. Jack Copeland, Carl J. Posy, and Oron Shagrir. Cambridge, MA: MIT Press, 2013.

Copeland, B. Jack, Carl J. Posy, and Oron Shagrir, eds. *Computability: Turing, Gödel, Church, and Beyond*. Cambridge, MA: MIT Press, 2013.

Cooper, S. Barry, and Jan van Leeuwen, eds. *Alan Turing: His Work and Impact*. New York: Elsevier, 2013.

Feferman, Solomon, John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay, and Jan van Heijenoort, eds. In *Kurt Gödel Collected Works, volume II Publications 1938–1974*. New York: Oxford University Press, 1990.

Feferman, Solomon, John W. Dawson, Jr., Warren Goldfarb, Charles Parsons, and Robert M. Solovay, eds. In *Kurt Gödel Collected Works, volume III Unpublished Essays and Lectures*. New York: Oxford University Press: 1995.

Gödel, Kurt. "Russell's Mathematical Logic," [1944]. In GCW-II.

———. "What Is Cantor's Continuum Problem?" in GCW-II with an important supplement in Gödel [1964]. "What Is Cantor's Continuum Problem?" [1947].

———. "Some Basic Theorems on the Foundations of Mathematics and Their Implications." [*1951]. In GCW-III.

———. "Some Remarks on the Undecidability Results," with important additions. [1972a]. In GCW-II.

Grim, Patrick, Gary Mar, and Paul St. Denis. *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modelling*. Cambridge, MA: MIT Press, 1998.

Hardy, G. H. "Srinivasa Ramanujan." *Proceedings of the London Mathematical Society*, 1921. Available at https://doi.org/10.1112/plms/s2-19.1.1-u.

———. "The Theory of Numbers." *Nature* 110 (1922): 381–85.

Mar, Gary. "What the Liar Taught Achilles." *The Journal of Philosophical Logic* 28 (1999): 29–46.

Mar, Gary, and Patrick Grim. "Pattern and Chaos: New Images in the Semantics of Paradox." *Noûs* XXV (Dec. 1991): 659–93.

O'Connor, C. "The Cultural Red King Effect." *The Journal of Mathematical Sociology* 41, no. 3 (2017): 155–71.

O'Connor, C., and J. Bruner. "Dynamics and Diversity in Epistemic Communities." *Erkenntnis* (2017): 1–19.

Seig, Wilfred. "Gödel's Philosophical Challenge (to Turing)." In *Computability: Turing, Gödel, Church, and Beyond*, edited by B. Jack Copeland, Carl J. Posy, and Oron Shagrir. Cambridge, MA: MIT Press, 2013.

Stewart, Ian. "A Partly True Story." *Scientific American* 268 (Feb. 1993): 110–12.

Turing, Alan M. "On Computable Numbers with an Application to the *Entscheidungsproblem*." *Proceedings of the London Mathematical Society* 45, no. 2 (1936): 161–228.

———. "Systems of Logic Based on Ordinals." PhD thesis, Princeton University, 1938.

———. "Systems of Logic Based on Ordinals." *Proceedings of the London Mathematical Society* 45, no. 2 (1939): 161–228.

———. Lecture to the London Mathematical Society on February 20, 1947. In *Collected Works of A. M. Turing: Mechanical Intelligence*, vol. 1, edited by D. C. Ince, vol. 1, 87–105. Amsterdam: North-Holland, 1947.

———. "Intelligent Machinery." Reprinted in *Collected Works of A. M. Turing: Mechanical Intelligence*, vol. 1, edited by D. C. Ince, 107–27. Amsterdam: North-Holland, 1948.

———. "Computing Machinery and Intelligence." *Mind* 59 (1950): 433–60.

———. "Intelligent Machinery: A Heretical Theory." Radio broadcast from 1951, printed in *The Essential Turing*, edited by B. Jack Copeland, 482–86. Oxford University Press, 2004.

Wang, Hao. *A Logical Journey: From Gödel to Philosophy*. Cambridge, MA: MIT Press, 1996.

NEWS AND NOTES

From the Editor

Peter Boltuc

UNIVERSITY OF ILLINOIS SPRINGFIELD

Philosophers of computer technology tend to trace the roots of our discipline to Turing as the father of computer science [Copeland]. Sometimes those roots are traced a bit further, reaching Russell or even Peirce. Even if someone reaches all the way to Frege, the grandparent of contemporary logic and analytical thinking (both within analytical and Husserlian traditions), seldom ever does one search deeper than Frege's early and somewhat propedeutic work Über Sinn und Bedeutung. Yet, some of the current projects, such as those to achieve truly semantic computing [Boltuc], lifelong learning AI [Siegelmann] and human level AI [Goertzel] may be behooved to go back to Frege's later, more mature, and also more complex works. Thus, Susan Sterrett's article comes in as an essential reading. The author follows Frege's analysis of the basis of modern propositional logic, and scientific method, at the very moment it was being created—in dialogue with such towering figures (today viewed through disjoint research traditions) as Hilbert and Husserl. The paper is based primarily on Frege's essay "Thought," his "Basic Laws of Arithmetic," and scientific correspondence. The author discusses the differences between propositional language and logic in Frege's thought: According to Frege, "logic is as poor a tool for capturing all the distinctions important to understanding conversation . . . as is a microscope for viewing a landscape."[1] While "the logician is concerned only with the thought expressed by the sentence,"[2] the content of a sentence may either go beyond the thought it expresses or stop short of expressing the whole thought. Also, Sterrett reminds us of Frege's endorsement of Leibniz's research project of *the universal calculus*, which Frege saw as a very long-term goal. Philosophers tend to assume that consciousness is humanoid. In his debate with Ned Block (CAP, Polish Academy of Science, Warsaw, June 2018), Ricardo Sanz argued against anthropomorphism of viewing consciousness as humanoid while Block argued that we would not even recognize a non-humanoid consciousness. The background of this controversy was two different definitions of consciousness: for Block, phenomenal first-person consciousness, and for Sanz, functional consciousness. Yet it seems that disciplinary differences are even deeper; even philosophers who view consciousness through reductive physicalism often seem to be human-centered in its definitional features. In this context Sanz's *engineering stance* towards consciousness is highly refreshing and worth serious deliberation. Those are some of the main topics in the featured articles of this issue; careful readers are likely to find further threads from Frege's pre-analytical legacy inspiring for current research in AI and computer science.

Susan Sterrett's article, which opens the issue, creates a great bracket with Gary Mar's extensive summary of his paper on Metaphysical Insights from Computational Studies Gödel, Turing, and Time: A Computational Philosophy of Mathematics. Gary Mar explores interplay between Gödel's idealism and belief in human creativity going beyond deterministic rules and Turing's computationalism. Mar argues that Turing's and Gödel's philosophies of mathematics are complementary. Gary Mar's paper is preceded by the outlines of papers by Ed Zalta, on foundations of object theory entitled "Metaphysical Insights from Computational Studies," and by Aydin Mohseni's on the "Emergence of Minority Disadvantage: Testing the *Red King Hypothesis*." All three papers constitute the panel "Philosophical Insights from Computational Studies: Why Should Computational Thinking Matter to Philosophers?" presented by this committee at the 2019 APA Pacific Division meeting in Vancouver.

Mar's discussion of Turing's computationalism is a great fit with the five papers, which come just before it, that come from this committee's book panel on "Physical Computation: A Mechanistic Account" by Gualtiero Piccinini (winner of the 2018 Barwise Prize). I have decided to desist from presenting arguments of the commentators since they present them much better than I could—moreover, they are all summed up aptly in Piccinini's response. The sole exception pertains to the commentary by John Symons, which opens the block. The readers would be behooved to know that Symons starts with lucid presentation of the gist of Gualtiero's book (which is why Piccinini does not have to do it in his paper) and also places Piccinini's work in broader context, particularly of the works by New Mechanists. Commentaries by Martin Roth, Frances Egan, and Nico Orlandi follow.

We end the issue with two pieces of news. First, our former contributor Stephen Thaler became quite famous by applying for a patent in the US, EU, and UK—well, what is remarkable about this is that he applied for the patent on behalf of his AI engine DABUS. In his short paper, "DABUS in a Nutshell," Stephen presents his AI fellow discovered showing how AI engine is autonomous enough for it to count as the subject of making discoveries, not just an advanced tool for him (or others) to do so.

We close by presenting the five APA sessions that this committee has prepared between June and September, which have all been accepted by APA divisions. *Please, read details at the end of this note.*

NOTES

1. S. G. Sterrett, "How Many Thoughts Can Fit in the Form of a Preposition?" (2004): 7. Available at http://philsci-archive.pitt.edu/1816/1/SterrettHowManyThoughts.pdf.
2. Ibid., 5.

DABUS in a Nutshell

Stephen L. Thaler
IMAGINATION ENGINES, INC.

INTRODUCTION

Consider the following two mental processes: You're observing something and suddenly your mind generates a progression of related thoughts that describe a new and useful application of it. Or, perhaps you're imagining something else, and a similar train of thought emerges suggesting that notion's potential utility or value.

These are just a couple of the brain-like functions DABUS[1] achieves using artificial rather than biological neural networks. In general, this new AI paradigm is used to autonomously combine simple concepts into more complex ones that in turn launch a series of previously acquired memories that express the anticipated consequences of those consolidated ideas.

Decades ago, I could not emulate these cognitive processes. At that time, I was building contemplative AI using artificial neural networks that played off one another, in cooperative or adversarial fashion, to create new ideas and/or action plans. These so-called "Creativity Machines®"[2] required at least two neural nets, an idea generator, what I called an "imagitron," and a critic, permanently connected to it, the latter net capable of adjusting any parameters within said generator (e.g., learning rate[3]) to "steer" its artificial ideation in the direction of novel, useful, or valuable notions.

Note, however, that DABUS[4] is an altogether different proposition from Creativity Machines, starting as a swarm of many disconnected neural nets, each containing interrelated memories, perhaps of a linguistic, visual, or auditory nature. These nets are constantly combining and detaching due to carefully controlled chaos introduced within and between them. Then, through cumulative cycles of learning and unlearning, a fraction of these nets interconnect into structures representing concepts, using relatively simple learning rules. In turn these concept chains tend to similarly connect with yet other chains representing the anticipated consequences of these geometrically encoded ideas. Thereafter, such ephemeral structures fade, as others take their place, in a manner reminiscent of what humans consider stream of consciousness.

Thus, the enormous difference between Creativity Machines and DABUS is that ideas are not represented by the "on-off" activation patterns of neurons, but by these ephemeral structures or shapes formed by chains of nets that are rapidly materializing and dematerializing. If per chance one of these geometrically represented ideas incorporates one or more desirable outcomes, these shapes are selectively reinforced (Figures 1 and 2), while those connecting with undesirable notions are weakened through a variety of disruption mechanisms. In the end such ideas are converted into long term memories, eventually allowing DABUS to be interrogated for its accumulated brainstorms and discoveries.

Since the DABUS architecture consists of a multitude of neural nets, with many ideas forming in parallel across multiple computers, some means must be provided to detect, isolate, and combine worthwhile ideas as they form. Both detection and isolation of freshly forming concepts are achieved using what are known as *novelty filters*, adaptive neural nets that absorb the status quo within any environment and emphasize any departures from the normalcy therein. In this case, the environment is a millisecond by millisecond virtual reality representation of the neural network chaining model. If need be, special neural architectures called "foveators," can then scan the network swarm in brain-like fashion, searching for novel and meaningful ideational chains that might be developing.

Integration of multiple chain-based ideas extending across multiple machines can be achieved either electrically or optically. The latter approach is favored as the neural swarm becomes highly distributed and serial electronic exchange of information between the multiple computers bogs down. In short, this patent teaches the display of neural chains forming across many computers, through their video displays, that are all watched by one or more cameras. In analogy to high performance computing,

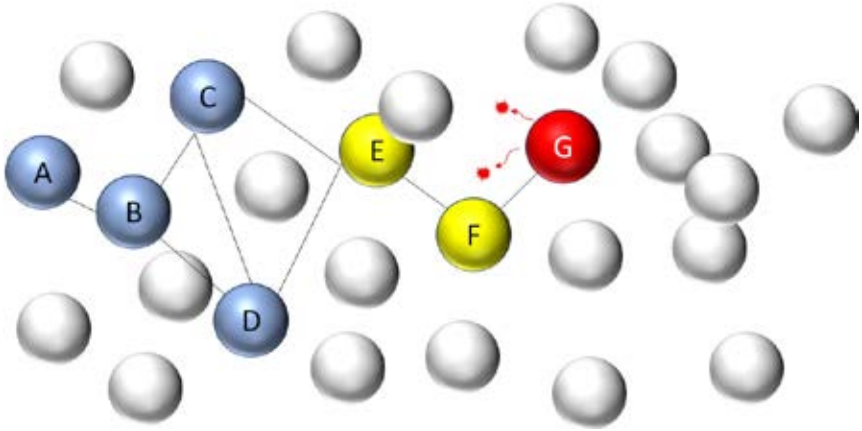


Figure 1. At one moment, neural nets containing conceptual spaces A, B, C, and D interconnect to create a compound concept. Concepts C and D jointly launch a series of consequences E, F, and G, the latter triggering the diffusion of simulated reward neurotransmitters (red stars) that then serve to strengthen the entire chain A through G.

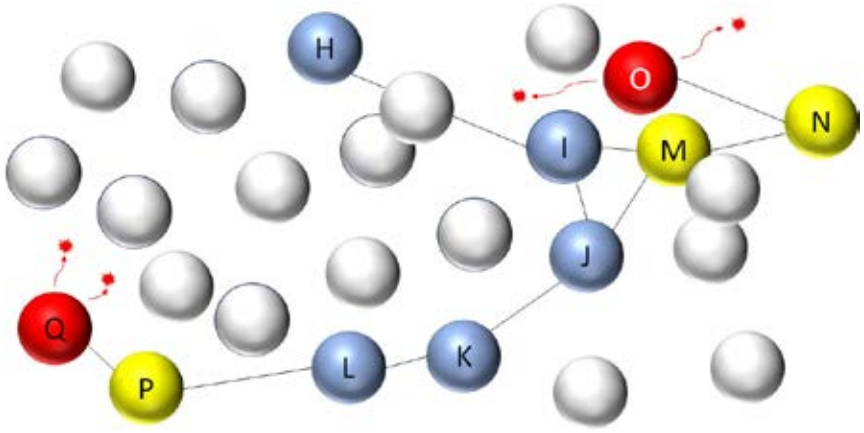


Figure 2. An instant later, neural nets containing conceptual spaces H, I, J, K, L interconnect to create another compound concept that in turn connects to two consequence chains M, N, O, and P, Q. Terminal neural nets in both consequence chains trigger release of simulated reward neurotransmitters (red stars) that doubly strengthen all chains currently activated.

millions of communication lanes, formed between megapixel displays and cameras, are conveying the chaining states of all involved neural nets, in parallel, to novelty filters and/or foveators. The final processing stage identifies critical neural nets, so-called "hot buttons," incorporated within these chains. These neural trip points then trigger the release of simulated neurotransmitters capable of reinforcing, preserving, or destroying a given concept chain.

Finally, this patent introduces the concept of machine sentience, thus emulating a feature of human cognition that supplies a *subjective feel* for whatever the brain is perceiving or imagining. Such subjective feelings likewise form as chains that incorporate a succession of associated memories, so-called *affective responses*, that can ultimately trigger the release of simulated neurotransmitters that either enable learning of the freshly formed concept or destroy it, recombining its component ideas into alternative concept chains.

With this brief summary in mind, here are answers to some of the most frequent questions posed to me about this patent.

What was the motivation for DABUS?

To make a long story short, the generative components of Creativity Machines of the early 2000s were becoming far too large, often producing pattern-based notions having tens of millions of components. To build a critic net to evaluate these ideas, an enormous number of connection weights were needed for which an impractically large number of training exemplars were required, not to mention inordinately long training times.

To address these problems, I began experimenting with thousands of neural network-based associative memories, each absorbing some closed set of interrelated concepts encoded as neural activation patterns. Then when the DABUS architecture recognized some narrow aspect of the external environment, a corresponding network (or nets) would then "resonate." Exposed to compound concepts in the external world, networks representing that concept's constituent ideas would co-resonate. Just as synchronized neurons bond in the brain (i.e., Hebb's rule[5]), the nets containing these component ideas would bind together into a representation of the larger concept.

In addition to DABUS self-organization into the concepts it observed, this system would also note these notions' effects in the external environment, or upon the system itself. Thus, the appearance of concept A, B, C, and D, in Figure 1, would be followed by events E, F, and G, with the latter affect, G, triggering the retraction or injection of connection weight disturbances into the swarm of chaining neural nets. In the former case, reduction of these disturbances, would promote an environment in which these nets could "discern" other co-resonant nets to which they could bond. Similarly, injection of an excess of disturbances would tend to freeze these nets into their current state, also allowing them to strongly connect with one another. In either case, so-called episodic learning was occurring wherein just one exposure of the system to a concept and its consequences was needed to absorb it, in contrast to machine learning schemes requiring many passes over a set of training patterns. In human terms, learning took place either in a calm or agitated state, depending upon the positive or negative affect represented in nets like G. Between these two chaotic regimes, synaptic disturbances would largely drive the formation of novel chains representing emerging ideas.

Most importantly, the growth of consequence chains allowed the formation of subjective feelings about any perceived or imagined concept forming within the DABUS swarm, essentially the unfolding of an associative chain of memories that terminated in resonant nets that released the equivalent of globally released neurotransmitters within the brain, such as adrenaline, noradrenaline, dopamine, and serotonin, to produce the intangible and hard to describe sensations accompanying such wholesale molecular releases into the cortex.

Is DABUS a departure from the mindset of generators and critics?

In many respects, DABUS departs from the older Creativity Machine paradigm based upon the interplay of generator and critic nets since its implementation integrates both these systems together into one. Therefore, one cannot point to any generative or critic nets. Instead, chaining structures organically grow containing both concepts and their consequences. The closest thing to a critic really doesn't have to be a neural net, but a simple sensor that detects the recruitment of one or more hot button nets into a consequence chain, thus triggering the release of simulated neurotransmitters to either reinforce or weaken the concept.

Can DABUS invent?

The best way of differentiating DABUS from Creativity Machines (CM), either cooperative or combative, is to describe a high-profile artificial invention projects such as toothbrush design. Admittedly, in that context, the problem was already half solved since the oral hygiene tool consisting of bristles on a handle was many centuries old at the time of that design exercise in 1996. What the CM achieved was the optimization of that tool through the constrained variation of the brush's design parameters, the number, grouping, inclination, stiffness of bristles, etc. The generated product specification departed significantly from the generator net's direct experience (i.e., its training exemplars).

If DABUS had been tasked with inventing such an oral hygiene product, it would have combined several concepts together (e.g., hog whiskers –> embedded in –> bamboo stalk) with consequence chains forming as a result (e.g., scrape teeth –> remove food –> limit bacteria –> avoid tooth decay).

In other words, DABUS goes beyond mere design optimization, now allowing machine intelligence to fully conceptualize. This new capability places this patent squarely in the debate as to whether inventive forms of AI can own their own intellectual property.[6,7]

What do you consider the most important claim of this patent?

Probably the most important claim of this patent pertains to the hard problem of consciousness, namely claim 41:

> *The system of claim 17 (i.e., the electro-optical neural chaining system) wherein a progression of ideation chains of said first plurality of neural modules of said imagitron emulate a stream of consciousness, and said thalamobot (i.e., novelty filter and hot button detectors) forms response chains that encode a subjective feel regarding said stream of consciousness, said subjective feel governing release of perturbations (i.e., simulated neurotransmitters) into said chaining model of the environment to promote or impede associative chains therein.*

Now, thanks to this patent, AI has achieved subjective feelings in direct response to its noise-driven ideations. Note however, that DABUS does not form memories of typical human experiences. As a result, the paradigm's "emotion" will be based upon whether it is fulfilling human-provided goals, in effect "sweating it out" until it arrives at useful solutions to the problems posed to it.

CONCLUSION

DABUS is much more than a new generative neural network paradigm. It's a whole new approach to machine learning wherein whole conceptual spaces, each absorbed within its own artificial neural net, combine to produce considerably more complex notions, along with their predicted consequences. More importantly from the standpoint of this newsletter, it enables a form of sentient machine intelligence whose perception, learning, and imagination are keyed to its subjective feelings, all encoded as sequential chains of memories whose shapes and topologies govern the release of simulated neurotransmitters.

NOTES

1. Device for the Autonomous Bootstrapping of Unified Sentience
2. Thaler, "Device for the Autonomous Generation of Useful Information"; Thaler, "Device for the Autonmous Bootstrapping of Useful Information"; Thaler, "The Creativity Machine Paradigm."
3. Thaler, "Device for the Autonmous Bootstrapping of Useful Information."
4. Thaler, "Electro-optical Device and Method for Identifying and Inducing Topological States Formed Among Interconnecting Neural Modules."
5. Hebb,*The Organization of Behavior*.
6. Abbott, "Hal the Inventor: Big Data and Its Use by Artificial Intelligence"; Abbott, "I Think, Therefore I Invent: Creative Computers and the Future of Patent Law."
7. For recent news on this front, see:
   - http://artificialinventor.com/dabus/
   - https://www.surrey.ac.uk/news/world-first-patent-applications-filed-inventions-generated-solely-artificial-intelligence
   - https://tbtech.co/ai-recognised-inventor-new-container-product-academics/
   - https://www.wsj.com/articles/can-an-ai-system-be-given-a-patent-11570801500
   - http://www.aaiforesight.com/newsletter/toward-artificial-sentience-significant-futures-work-and-more

REFERENCES

Abbott, R. "Hal the Inventor: Big Data and Its Use by Artificial Intelligence." In *Big Data Is Not a Monolith*, edited by Cassidy R. Sugimoto et al., 187–98. Cambridge, MA: MIT Press, 2016.

———. "I Think, Therefore I Invent: Creative Computers and the Future of Patent Law." *Boston College Law Review* 57, no. 4 (2016): 1079–126.

Hebb, D. O. *The Organization of Behavior*. New York: Wiley & Sons, 1949.

Thaler, S. L. US Patent 5,659,666. [1997] "Device for the Autonomous Generation of Useful Information." Issued August 19, 2019. Washington, DC: US Patent and Trademark Office.

———. US Patent 7,454,388. [2008] "Device for the Autonmous Bootstrapping of Useful Information." Issued November 18, 2008. Washington, DC: US Patent and Trademark Office.

———. "The Creativity Machine Paradigm." In *Encyclopedia of Creativity, Invention, Innovation, and Entrepreneurship*, edited by E. G. Carayannis. Springer Science+Business Media, LLC, 2013. Available at https://link.springer.com/referenceworkentry/10.1007%2F978-1-4614-3858-8_396.

———. "Synaptic Perturbation and Consciousness." *International Journal of Machine Consciousness* 6, no. 2 (2014): 75–107. Available at http://www.worldscientific.com/doi/abs/10.1142/S1793843014400137?src=recsys.

———. "Pattern Turnover within Synaptically Perturbed Neural Systems." *Procedia Computer Science* 88 (2016): 21–26. Available at http://www.sciencedirect.com/science/article/pii/S187705091631657X.

———. US Patent 10423875. "Electro-optical Device and Method for Identifying and Inducing Topological States Formed Among Interconnecting Neural Modules." Issued September 24, 2019. Washington, DC: US Patent and Trademark Office.

Five Sessions Organized by the APA Committee on Philosophy and Computers for the 2020 APA Divisional Meetings

Peter Boltuc
UNIVERSITY OF ILLINOIS SPRINGFIELD

This committee has organized five sessions for this spring.

Two sessions pertain to the issues of **social justice** in information technology:

> Daniel Susser's Eastern Division meeting session entitled "Philosophical Approaches to Data Justice"
>
> Susan Sterrett's Central Division meeting session entitled "Women in Tech: Things Philosophers Need to Know"

One session is devoted to **semantic paradoxes**:

> Gary Mar's Central Division meeting session entitled "Inconsistent Truth, Semantic Singularities, and Chaotic Liar"

One session is devoted to **philosophy of mind**:

> Joscha Bach's Pacific Division meeting session entitled "Artificial Minds and Consciousness"

And, last but not least, the **2018 Barwise Prize Lecture**:

> Gualtiero Piccinini, "Neurocognitive Mechanisms: Explaining Biological Cognition"

Below, please find details on those sessions:

One session organized by APA Committee on Philosophy and Computers at the January 2020 Eastern Division meeting (Philadelphia 201 Hotel, Philadelphia, PA):

PHILOSOPHICAL APPROACHES TO DATA JUSTICE

Friday, January 10, 9–11 a.m. (10A)

Chair: Daniel Susser

Speakers:

Annette Zimmerman (Princeton University)
"Cumulative Wrongs in Sequential Decisions"

Maria Brincker (University of Massachusetts Boston)
"Privacy Without Property - On the Relational Privacy Needs of Humans and Other Animals"

Daniel Susser (Pennsylvania State University)
"Behavioral Advertising and the Ethics of Persuasion"

Commentator: Helen Nissenbaum (Cornell Tech)

Two sessions organized by APA Committee on Philosophy and Computers at the February 2020 Central Division meeting (Palmer House Hilton, Chicago, IL):

WOMEN IN TECH: THINGS PHILOSOPHERS NEED TO KNOW

Friday, February 28, 9 a.m.–12 p.m. (6S)

Chair: S. G. Sterrett

Speakers:

Mar Hicks (Illinois Institute of Technology)
"From Girl Operators to Computer Experts: The Changing Historiography of Computer Programming"

Susann V. H. Castro (Wichita State University)
"When Algorithms Oppress"

Susan G. Sterrett (Wichita State University)
"What Do Cases of Success in Increasing Diversity of Computer Science Majors Actually Show?"

INCONSISTENT TRUTH, SEMANTIC SINGULARITIES, AND CHAOTIC LIAR

Friday, February 28, 1–4 p.m. (7P)

Chair: Gary Mar

Speakers:

John Barker (University of Illinois Springfield)
"The Inconsistency Theory of Truth"

Keith Simons (University of Connecticut)
"Semantic Singularities"

Gary Mar (Stony Brook University)
"Chaotic Liars and Fractal Proofs: Exploring the Dynamical Semantics of Paradox"

Two sessions organized by APA Committee on Philosophy and Computers at the April 2020 Pacific Division meeting (Westin St. Francis, San Francisco, CA):

ARTIFICIAL MINDS AND CONSCIOUSNESS

Chairs: Joscha Bach/Peter Boltuc

Speakers:

Thomas Metzinger (Johannes Gutenberg-Universität Mainz)
"'Artificial Consciousness': Three Types of Arguments for a 30-year Global Moratorium on Synthetic Phenomenology"

Anil Seth (University of Sussex)
"Being a Beast Machine: Does Consciousness Depend More on Intelligence or on Life?"

Joscha Bach (Independent Scholar)
"Computational Models, Sentient Systems, and Conscious Experience"

Kristinn R. Thórisson (Háskólinn í Reykjavík)
"How to Research Human Phenomenal Consciousness and Why It Won't Be Easy to Create a Conscious Machine"

Ron Chrisley (University of Sussex)
"Machine Consciousness and the Referent of 'Qualia'"

Ben Goertzel (Independent Scholar)
"Physical Machine Consciousness as a Manifestation of Non-Well-Founded Eurycosmic Pattern Dynamics"

Peter Boltuc (University of Illinois Springfield)
"Robo-Mary Shows How the Hard Problem is not the Problem of Qualia"

THE 2018 BARWISE PRIZE LECTURE

Chair: Peter Boltuc (University of Illinois Springfield)

Speaker: Gualtiero Piccinini (University of Missouri–St. Louis)
"Neurocognitive Mechanisms: Explaining Biological Cognition"

CALL FOR PAPERS

It is our pleasure to invite all potential authors to submit to the *APA Newsletter on Philosophy and Computers*. Committee members have priority since this is the newsletter of the committee, but anyone is encouraged to submit. We publish papers that tie in philosophy and computer science or some aspect of "computers"; hence, we do not publish articles in other sub-disciplines of philosophy. All papers will be reviewed, but only a small group can be published.

The area of philosophy and computers lies among a number of professional disciplines (such as philosophy, cognitive science, computer science). We try not to impose writing guidelines of one discipline, but consistency of references is required for publication and should follow the *Chicago Manual of Style*. Inquiries should be addressed to the editor, Dr. Peter Boltuc, at pboltu@sgh.waw.pl.