NEWSLETTER | The American Philosophical Association

Philosophy and Computers

SPRING 2019

VOLUME 18 | NUMBER 2

VOLUME 18 | NUMBER 2

SPRING 2019

ISSN 2155-9708

APA NEWSLETTER ON

Philosophy and Computers

PETER BOLTUC, EDITOR

VOLUME 18 | NUMBER 2 | SPRING 2019

FEATURED ARTICLE

Turing's Mystery Machine

Jack Copeland and Diane Proudfoot

UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NZ

ABSTRACT

This is a detective story. The starting-point is a philosophical discussion in 1949, where Alan Turing mentioned a machine whose program, he said, would in practice be "impossible to find." Turing used his unbreakable machine example to defeat an argument against the possibility of artificial intelligence. Yet he gave few clues as to how the program worked. What was its structure such that it could defy analysis for (he said) "a thousand years"? Our suggestion is that the program simulated a type of cipher device, and was perhaps connected to Turing's postwar work for GCHQ (the UK equivalent of the NSA). We also investigate the machine's implications for current brain simulation projects.

INTRODUCTION

In the notetaker's record of a 1949 discussion at Manchester University, Alan Turing is reported as making the intriguing claim that—in certain circumstances—"it would be impossible to find the programme inserted into quite a simple machine."[1] That is to say, reverse-engineering the program from the machine's behavior is in practice not possible for the machine and program Turing was considering.

This discussion involved Michael Polanyi, Dorothy Emmet, Max Newman, Geoffrey Jefferson, J.Z. Young, and others (the notetaker was the philosopher Wolfe Mays). At that point in the discussion, Turing was responding to Polanyi's assertion that "a machine is fully specifiable, while a mind is not." The mind is "only said to be unspecifiable because it has *not yet been* specified," Turing replied; and it does not follow from this, he said, that "the mind is unspecifiable"—any more than it follows from the inability of investigators to specify the program in Turing's "simple machine" that this program is unspecifiable. After all, Turing knew the program's specification.

Polanyi's assertion is not unfamiliar; other philosophers and scientists make claims in a similar spirit. Recent examples are "mysterianist" philosophers of mind, who claim that the mind is "an ultimate mystery, a mystery that human intelligence will never unravel."[2] So what was Turing's machine, such that it might counterexample a claim like Polanyi's? A machine that—although "quite a simple" one—thwarted attempts to analyze it?

A "SIMPLE MACHINE"

Turing again mentioned a simple machine with an undiscoverable program in his 1950 article "Computing Machinery and Intelligence" (published in *Mind*). He was arguing against the proposition that "given a discrete-state machine it should certainly be possible to discover by observation sufficient about it to predict its future behaviour, and this within a reasonable time, say a thousand years."[3] This "does not seem to be the case," he said, and he went on to describe a counterexample:

> I have set up on the Manchester computer a small programme using only 1000 units of storage, whereby the machine supplied with one sixteen figure number replies with another within two seconds. I would defy anyone to learn from these replies sufficient about the programme to be able to predict any replies to untried values.[4]

These passages occur in a short section titled "The Argument from Informality of Behaviour," in which Turing's aim was to refute an argument purporting to show that "we cannot be machines."[5] The argument, as Turing explained it, is this:

> (1) If each man had a definite set of laws of behaviour which regulate his life, he would be no better than a machine.
>
> (2) But there are no such laws.
>
> ∴ (3) Men cannot be machines.[6]

Turing agreed that "being regulated by laws of behaviour implies being some sort of machine (though not necessarily a discrete-state machine)," and that "conversely being such a machine implies being regulated by such laws."[7] If this biconditional serves as a reformulation of the argument's first premiss, then the argument is plainly valid.

Turing's strategy was to challenge the argument's second premiss. He said:

> we cannot so easily convince ourselves of the absence of complete laws of behaviour . . . The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say "We have searched enough. There are no such laws."[8]

Turing then offered his example of the discrete-state machine that cannot be reverse-engineered, to demonstrate "more forcibly" that the failure to find laws of behavior does not imply that no such laws are in operation.[9]

These are the only appearances of Turing's "simple machine" in the historical record (at any rate, in the declassified record). How could Turing's mysterious machine have worked, such that in practice it defied analysis? And what implications might the machine have for brain science and the philosophy of mind—beyond Turing's uses of the machine against Polanyi's bold assertion and against the "informality of behaviour" argument? We discuss these questions in turn.

One glaringly obvious point about Turing's mystery machine (henceforward "MM") is that it amply meets the specifications for a high-grade cipher machine. It is seldom noted that Turing's career as a cryptographer did not end with the defeat of Hitler. During the post-war years, as well as playing a leading role in Manchester University's Computing Machine Laboratory, Turing was working as a consultant for GCHQ, Bletchley Park's peacetime successor.[10] With the development of the first all-purpose electronic computers, two of Turing's great passions, computing and cryptography, were coalescing. He was an early pioneer in the application of electronic stored-program computers to cryptography.

The Manchester computer's role in Cold War cryptography remains largely classified. We know, however, that while the computer was at the design stage, Turing and his Manchester colleague Max Newman—both had worked on breaking the German "Tunny" cipher system at Bletchley Park—directed the engineers to include special facilities for cryptological work.[11] These included operations for differencing (now a familiar cryptological technique, differencing originated in Turing's wartime attack on the Tunny cipher system, and was known at Bletchley Park as "delta-ing"). GCHQ took a keen interest in the Manchester computer. Jack Good, who in 1947 had a hand in the design of Manchester's prototype "Baby" computer, joined GCHQ full-time in 1948.[12] Others at Manchester who were closely involved with the computer also consulted for GCHQ;[13] and a contingent from GCHQ attended the inaugural celebration for what Turing called the Mark II[14] version of the Manchester computer, installed in Turing's lab in 1951. The question of how to program electronic digital computers to encrypt military and commercial material was as new as it was promising. GCHQ installed a Mark II in its new headquarters at Cheltenham.[15]

MM AS AN ENCRYPTION DEVICE

How might MM be used as a cipher machine? A hypothetical example will illustrate the general principles. Suppose Alice wishes to encipher her message "I LUV U" (the "plaintext") before sending the result (the "ciphertext") to Bob. Bob, who knows Alice's enciphering method, will uncover the plaintext by using Alice's method in reverse.

Alice's first step is to convert the plaintext into binary. Turing would have done this using teleprinter code (also known as Baudot-Murray code). Employed worldwide in communications systems at that time, teleprinter code transformed each keyboard character into a different string of five bits; for example, *A* was 11000 and *B* was 10011. Teleprinter code is the ancestor of the ASCII and UTF-8 codes used today to represent text digitally. Turing was very familiar with teleprinter code from his time at Bletchley Park, since the German Tunny system used it. In fact, Turing liked teleprinter code so much that he chose it as the basis for the Manchester computer's programming language.

To convert the plaintext into binary, Alice needs to know the following teleprinter code equivalences: "I" is 01101; "L" is 01001; "U" is 11100; "V" is 01111; and space is 00100. To do the conversion, she first writes down the teleprinter code equivalent of "I," and then (writing from left to right) the teleprinter code equivalent of space, and then of "L," and so on, producing:

01101001000100111100011110010011100

This string of 35 figures (or bits) is called the "binary plaintext."

So far, there has been no encryption, only preparation. The encryption will be done by MM. Recall that MM takes a sixteen-figure number as input and responds with another sixteen-figure number. Alice readies the binary plaintext for encryption by splitting it into two blocks of sixteen figures, with three figures "left over" on the right:

```
0110100100010011     1100011110010011     100
```

Next, she pads out the three left-over figures so as to make a third sixteen-figure block. To do this, she first adds "/" (00000), twice, at the end of the binary plaintext, so swelling the third block to thirteen figures, and then she adds (again on the far right of the third block) three more bits, which she selects at random (say 110), so taking the number of figures in the third block to sixteen. The resulting three blocks form the "padded binary plaintext":

```
0110100100010011     1100011110010011     1000000000000110
```

Alice now uses MM to encrypt the padded binary plaintext. She inputs the left-hand sixteen-figure block and writes down MM's sixteen-figure response; these are the first sixteen figures of the ciphertext. Then she inputs the middle block, producing the next sixteen figures of the ciphertext, and then the third block. Finally, she sends the ciphertext, forty-eight figures long, to Bob. Bob splits up the forty-eight figures of ciphertext into three sixteen-figure blocks and decrypts each block using his own MM (set up identically to Alice's); and then, working from the left, he replaces the ensuing five-figure groups with their teleprinter code equivalent characters. He knows to discard any terminal occurrences of "/", and also any group of fewer than five figures following the trailing "/". Bob is now in possession of Alice's plaintext.

This example illustrates how MM could have been used for cryptography; it gets us no closer, however, to knowing how MM generated its sixteen-figure output from its input. Probably this will never be known—unless the classified historical record happens to include information about MM's program, which seems unlikely. But let us speculate. The leading cipher machines of that era—Enigma, Tunny, the Hagelin, the British Typex and Portex, and Japanese machines such as Purple—all used a system of code-wheels to produce the ciphertext from the plaintext. We shall focus on Tunny, since it is the simplest of these machines to describe, and also because of its importance: the method of encryption pioneered in Tunny was a staple of military and commercial cryptosystems for many decades after the war. At Bletchley Park, Turing had invented the first systematic method for breaking the German Army's Tunny messages; it is quite possible that he was interested after the war in refining the machine's principles of encryption for future applications.

SIMULATING CODE-WHEEL MACHINES

The Tunny machine had at its heart twelve code-wheels,[16] but here we shall focus on a form of the Tunny machine with only ten code-wheels. Turing's wartime Tunny-breaking colleagues Jack Good and Donald Michie have argued persuasively that if (counterfactually) the Germans had used this ten-wheel version of the machine, it would have offered a far higher level of crypto-security than the twelve-wheel machine.[17] In fact, Michie remarked that, had the Germans used the ten-wheel version, "it is overwhelmingly probable that Tunny would never have been broken." With the ten-wheel machine, he said, there would be no "practical possibility of reverse-engineering the mechanism that generated it."[18] Assuming that the machine was not compromised by security errors, and the state of the art in cryptanalysis persisted much as it was in 1949, then the ten-wheel Tunny might indeed have remained unbroken for Turing's "a thousand years." If Turing was interested in Tunny post-war, it was most probably in this form of the machine.

As far as the user is concerned, the Tunny machine (both the ten- and twelve-wheel versions) is functionally similar to MM. When supplied with one five-figure number, the Tunny machine responds with another. When the number that is supplied (either by keyboard or from punched paper tape) is the teleprinter code of a letter of plaintext, the machine's reply provides the corresponding five figures of ciphertext. If, on the other hand, the machine is being used, not to encrypt the plaintext, but to decrypt the ciphertext, then its reply to five figures of ciphertext is the teleprinter code of the corresponding plaintext letter.

The machine produces its reply by first generating five figures internally, and then "adding" these to the number that is supplied as input. Tunny "addition" is better known to logicians as exclusive disjunction: 0 + 0 = 0, 1 + 0 = 1, 0 + 1 = 1, and 1 + 1 = 0. For example, if the incoming five figures are 01101, and the internally generated five figures are 00100, then the machine's reply is 01001 (i.e., 01101 + 00100).

The function of the code-wheels is to generate the five figures that are added to the incoming number. A simple way to generate five figures is to use an arrangement of five wheels, each of which contributes one figure. However, the setup actually used in the twelve-wheel Tunny machine (and the same in the ten-wheel version) is more complicated, the aim being greater security. Rather than a single group of five wheels, there are two groups, with five wheels in each group. In Bletchley Park jargon, the two groups were known respectively as the "X-wheels" and the "Ψ-wheels." Each group of wheels produces five figures, and these two five-figure numbers are then added together. It is the result of this addition that the machine goes on to add to the incoming number.

The Tunny machine's action is described by the machine's so-called "encipherment equation":

$$(\mathrm{X} + \Psi) + \mathrm{P} = \mathrm{C}$$

Adding the number X that is produced by the X-wheels to the number Ψ produced by the Ψ-wheels, and then adding the resulting number to P—the incoming five figures of binary plaintext—produces C, the corresponding five figures of ciphertext. With each incoming five-figure number, every wheel of the 10-wheel machine turns forwards a step; this has the result that the internally-generated number $\mathrm{X} + \Psi$ is always changing. (Incidentally, the function of the twelve-wheel Tunny's two extra wheels was quite different. These, known as the "motor wheels," served to create irregularities in the motions of the Ψ-wheels. No doubt the engineers at Lorenz[19] thought this arrangement would enhance the security of the machine, but they were badly mistaken. The motor wheels introduced a serious weakness, and this became the basis of Bletchley Park's highly successful attack on the twelve-wheel Tunny machine.)

One last relevant detail about Tunny's wheels. Each wheel had pins spaced regularly around its circumference. An operator could set each pin into one of two different positions, protruding or not protruding. (For security, the positions were modified daily.[20]) An electrical contact read figures from the rotating wheel (one contact per wheel): a pin in the protruding position would touch the contact, producing 1 (represented by electricity flowing), while a non-protruding pin would miss the contact, producing 0 (no flow). As a group of five wheels stepped round, the row of five contacts delivered five-figure numbers. Each wheel had a different number of pins, ranging from 23 to 61; at Bletchley Park, this number was referred to as the "length" of the wheel.

It would have been completely obvious to the post-war pioneers of computerized cryptography that one way to create a secure enciphering program was to simulate an existing secure machine. Turing's mystery machine may well have been a simulation of the ten-wheel Tunny machine, or of some other wheeled cipher machine.

Turing said that MM required "1000 units of storage." In the Manchester computer as it was in 1949–1950, a unit of high-speed storage consisted of a line of 40 bits spread horizontally across the screen of a Williams tube.[21] (A Williams tube, the basis of the computer's high-speed memory, was a cathode ray tube; a small dot of light on the tube's screen represented 1 and a large dot 0.) 1000 units is therefore 40,000 bits of storage. To simulate the ten-wheel Tunny on the Manchester computer, Turing would have needed ten variable-length shift registers to represent the wheels. Since the lengths of the ten wheels were, respectively, 41, 31, 29, 26, 23, 43, 47, 51, 53, and 59, a total of 403 bits of storage would be required for the pin patterns. This leaves more than 39 kilobits, an ample amount for storing the instructions—which add X, Ψ and P, shift the bits in the wheel registers (simulating rotation), and perform sundry control functions—and for executing them. Turing gave in effect an upper bound on the number of instruction-executions that occurred in MM in the course of encrypting one sixteen-figure number: MM gives its reply "within two seconds," he said. In 1949–1950, most of the Manchester computer's instructions took 1.8 milliseconds to execute; so approximately 1000 instructions could be implemented in two seconds.

Why are the numbers encrypted by MM sixteen figures long? This might indicate that MM simulated a machine with more than ten wheels. Or possibly a Tunny with modifications introduced by Turing for greater security; he might have increased the number of X-wheels and Ψ-wheels (and also the lengths of the wheels), or made other modifications that are impossible now to reconstruct. However, the number sixteen might in fact be no guide at all to the number of wheels. During 1941, when Tunny was first used for military traffic, German operating procedures made it transparent that the new machine had twelve wheels—invaluable information for the British cryptanalysts. Turing's choice of sixteen-figure numbers (rather than some number of figures bearing an immediate relationship to the number of wheels) might simply have been a way of masking the number of wheels.

Our first question about Turing's mystery machine was: How could it have worked, such that in practice it defied analysis? A not unlikely answer is: by simulating a ten-wheel Tunny or other Tunny-like machine. We turn now to our second question: Does MM have implications for brain science and the philosophy of mind, over and above Turing's uses of it against the argument from informality of behaviour and against the claim that "a machine is fully specifiable, while the mind is not"?

MM AND BRAIN SIMULATION

According to an incipient meme going by the name "Turing's Wager" (which has entries in Wikipedia and WikiVisually, as well as a YouTube video "Turing's Wager – Know It ALL"), the answer to our second question is a resounding *yes*.[22]

The term "Turing's Wager" seems to have been introduced in a 2017 journal article, "The Difficult Legacy of Turing's Wager," by Andrew Thwaites, Andrew Soltan, Eric Wieser, and Ian Nimmo-Smith. This article says:

> Turing introduced . . . *Turing's Wager* in . . . "Computing Machinery and Intelligence" . . . Turing's Wager (as we refer to it here) is an argument aiming to demonstrate that characterising the brain in mathematical terms will take over a thousand years.[23]

According to Thwaites et al., Turing viewed the project of describing "the human brain in mathematical terms" with "blunt scepticism."[24] They continue:

> [Turing] sought to highlight the challenges involved with a practical illustration . . . by writing a short computer program on his departmental workstation at the University of Manchester. This program accepted a single number, performed a series of unspecified calculations on it, and returned a second number. It would be extremely difficult, Turing argued, for anyone to guess these calculations from the input and output numbers alone. Determining the calculations taking place in the brain, he reasoned, must be harder still: not only does the brain accept tens-of-thousands of inputs from sensory receptors around the body, but the calculations these inputs undergo are far more complicated than anything written by a single programmer. Turing underscored his argument with a wager: that it would take an investigator at least a thousand years to guess the full set of calculations his Manchester program employed. Guessing the full set of calculations taking place in the brain, he noted, would appear prohibitively time-consuming (Turing 1950).[25]

However, there is no argument in "Computing Machinery and Intelligence" (nor elsewhere in Turing's writings) aiming to demonstrate that "characterising the brain in mathematical terms will take over a thousand years." The only conclusion that Turing drew from the MM example was (as described above) that failing to find the laws of behaviour or a full specification does not imply that none exist. It is false that "he noted" anything to the effect that "[g]uessing the full set of calculations taking place in the brain would appear prohibitively time-consuming," or that he "reasoned" in "Computing Machinery and Intelligence" about the difficulty of determining "the calculations taking place in the brain." Thwaites et al. tell us that Turing was not "optimistic about [the] chances of beating Turing's Wager,"[26] but this is an extraordinary claim—he never mentioned the so-called Wager.

On the other hand, perhaps the fact that Turing did not state or suggest Turing's Wager is of only historical or scholarly importance. If valid, the wager argument is certainly significant, since—as Thwaites et al. emphasize—it has important implications for the feasibility of current ambitious brain-modelling projects, such as the *BRAIN Initiative* in the United States, the European *Human Brain Project*, Japan's *Brain/MINDS Project*, and the *China Brain Project*. The wager argument, it is said, claims no less than that "it is impossible to infer or deduce a detailed mathematical model of the human brain within a reasonable timescale, and thus impossible in any practical sense."[27]

But is the argument valid? Set out explicitly, the wager argument is as follows:

(1) It would take at least 1,000 years to determine the calculations occurring in MM.

(2) The calculations occurring in the brain are far more complicated than those occurring in MM.

∴ (3) It would take well over 1,000 years to determine the calculatios occurring in the brain.

Both (1) and (2) are true, we may assume (certainly the calculations done by the ten-wheel Tunny are extremely simple in comparison with those taking place in the brain). However, these premises do *not* entail (3). If MM is a cryptographic machine, carefully and cleverly designed to thwart any efforts to determine the calculations taking place within it, there is no reason why a more complicated but potentially more transparent machine should not succumb to analysis more quickly than MM. The mere possibility that MM is a secure crypto-machine, impenetrable by design, shows that in some possible world (1), (2), and the negation of (3) are true, and thus that the "Turing's wager" argument is invalid. Unsurprising, therefore, that Turing did not offer the argument.

The answer to our second question, then, is *no*: MM has nothing to tell us about the prospects of brain-simulation.

CONCLUSION

In the 1949 Manchester discussion, Turing employed one of his hallmark techniques: attacking a grand thesis with a concrete counterexample. He used MM to undermine both Polanyi's claim that "a machine is fully specifiable, while a mind is not" and the "Informality of Behaviour" argument against artificial intelligence. However, as we argued, MM cannot further be used to undermine the—admittedly quite optimistic—claims proffered on behalf of large-scale brain simulation projects.

Turing himself made no connection between MM and the prospects for brain-simulation. One may still ask, though: What might Turing have thought of the *BRAIN Initiative* and other large-scale brain-modelling projects? It is impossible to say—but Turing was, after all, an early pioneer of brain-modelling. Not long after the war, he wrote:

> In working on the ACE I am more interested in the possibility of producing models of the action of the brain than in the practical applications to computing. . . . [A]lthough the brain may in fact operate by changing its neuron circuits by the growth of axons and dendrites, we could nevertheless make a model, within the ACE, in which this possibility was allowed for, but in which the actual construction of the ACE did not alter.[28]

Turing might well have cheered on his twenty-first-century descendants.

NOTES

[1] Copeland, ed. "'The Mind and the Computing Machine', by Alan Turing and Others."

[2] McGinn, *The Mysterious Flame: Conscious Minds in a Material World*, 5. McGinn is here describing not only the "mind," but the "bond between the mind and the brain."

[3] Turing, "Computing Machinery and Intelligence," 457.

[4] Ibid.

[5] Ibid.

[6] Turing first stated the argument in this form:

> "If each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines." ("Computing Machinery and Intelligence," 457)

He then considered the argument that results if "we substitute 'laws of behaviour which regulate his life' for 'laws of conduct by which he regulates his life'" (ibid.).

[7] Ibid., 457.

[8] Ibid.

[9] Ibid.

[10] Copeland "Crime and Punishment," 37.

[11] Tom Kilburn in interview with Copeland, July 1997; G. C. Tootill, "Informal Report on the Design of the Ferranti Mark I Computing Machine" (November 1949): 1 (National Archive for the History of Computing, University of Manchester).

[12] Copeland, "The Manchester Computer: A Revised History," 5–6, 28–29.

[13] Ibid., 6.

[14] The computer that Turing called the Mark II is also known as the Ferranti Mark I, after the Manchester engineering firm that built it.

[15] The manufacturer's name for the model installed at GCHQ was the Ferranti Mark I Star.

[16] Copeland, "The German Tunny Machine."

[17] Good and Michie, "Motorless Tunny."

[18] Ibid., 409.

[19] The Tunny machine was manufactured by the Berlin engineering company C. Lorenz AG, and for that reason was also called the "Lorenz machine" at post-war GCHQ (although never at wartime Bletchley Park, where the manufacturer was unknown and the British codename "Tunny machine" was invariably used).

[20] From August 1, 1944.

[21] Turing described the computer as it was at that time in an Appendix to Turing, *Programmers' Handbook for Manchester Electronic Computer Mark II*, entitled "The Pilot Machine (Manchester Computer Mark I)."

[22] See https://en.wikipedia.org/wiki/Turing%27s_Wager; and https://wikivisually.com/wiki/Turing%27s_Wager; https://www.youtube.com/watch?v=ONxwksicpV8.

[23] Thwaites et al., "The Difficult Legacy of Turing's Wager," 3.

[24] Ibid., 1.

[25] Ibid., 1–2.

[26] Ibid., 3.

[27] See https://en.wikipedia.org/wiki/Turing%27s_Wager.

[28] Letter from Turing to W. Ross Ashby, no date, but before October 1947. (Woodger Papers, Science Museum, London, catalogue reference M11/99). In *The Essential Turing*, 375.

REFERENCES

Copeland, B. J. "The German Tunny Machine." In Copeland et al. 2006.

———. "The Manchester Computer: A Revised History." *IEEE Annals of the History of Computing* 33 (2011): 4–37.

———. "Crime and Punishment." In Copeland et al. 2017.

Copeland, B. J., ed. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life*. Oxford & New York: Oxford University Press, 2004.

Copeland, B. J., ed. "'The Mind and the Computing Machine', by Alan Turing and Others." *The Rutherford Journal: The New Zealand Journal for the History and Philosophy of Science and Technology* 1 (2005). Available at http://www.rutherfordjournal.org/article010111.html.

Copeland, B. J., et al. *Colossus: The Secrets of Bletchley Park's Codebreaking Computers*. Oxford & New York: Oxford University Press, 2006.

Copeland, B. J., J. Bowen, M. Sprevak, R. Wilson, et al. *The Turing Guide*. Oxford & New York: Oxford University Press, 2017.

Good, I. J., and D. Michie. "Motorless Tunny." In Copeland et al. 2006.

McGinn, C. *The Mysterious Flame: Conscious Minds in a Material World*. New York: Basic Books, 1999.

Thwaites, A., A. Soltan, E. Wieser, and I. Nimmo-Smith. "The Difficult Legacy of Turing's Wager." *Journal of Computational Neuroscience* 43 (2017): 1–4.

Turing, A. M. *Programmers' Handbook for Manchester Electronic Computer Mark II*. Computing Machine Laboratory, University of Manchester; no date, circa 1950. http://www.alanturing.net/turing_archive/archive/m/m01/M01-001.html

Turing, A. M. "Computing Machinery and Intelligence." In *The Essential Turing*.

ARTICLES

Systems with "Subjective Feelings": The Logic of Conscious Machines

Igor Aleksander
IMPERIAL COLLEGE, LONDON

INTRODUCTION

> These are Christof Koch's closing remarks at the 2001 Swartz Foundation workshop on Machine Consciousness, Cold Spring Harbour Laboratories:
>
> ". . . we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artefacts designed or evolved by humans."

This account is aimed at identifying a formal expression of the "subjective feelings in artefacts" that Koch saw as being central to the definition of a conscious machine. It is useful to elaborate "artefacts" as the set of systems that have a physically realizable character and an analytic description. A "basic guess," first suggested in 1996,[1] and used since then, governs the progress of this paper: that the acceptedly problematic mind-brain relationship may be found and analyzed in the operation of a specific class of neural, experience-building machines. The paper is a journey through a progressive refinement of the characteristics of such machines.

The desired set of machines, (M,F) is characterized by having inner state structures that encompass subjective feelings (M for "machine," F for "feelings"). Such inner state structures are subjective for encompassing up-to-the-point, lifetime, external influences on the machine constituting (as will be argued) the mental experience of the machine. It is further argued that such state structures are available to the machine to determine future action or no-action deliberation. It is of interest that, equally stimulated by Koch's challenge, Franklin, Baars, and Ramamurthy[2] indicated that a stable, coherent perceptual field would add phenomenality to structures such as "Global Workspace." This stable field is a state of the broader concept of state structures that occur later in this paper. Also, since the Franklin paper, it has become problematic to discuss consciousness solely in terms of perceptual fields. As discussed below under "Cognitive Phenomenology," there is a need to discuss consciousness as it occurs in more abstract notions.

To start as generally as possible, M is taken to be the set of all machines that can have formal descriptions, that is, the set of all "artefacts designed or evolved by humans." The aim of this account is to develop a logical description of the set $M(F) \subset M$.

This is based, first, on satisfying a logical requirement that the internal states of M(F) be phenomenological (that is, be *about* the surrounding world and there being something it is like to be in such states), second, to define a logical structure which leads to such states becoming the subjective inner states of the artefact, third, how such subjectivity becomes structured into an inner state structure that is a formal candidate for the "conscious mind" of the individual artefact, and, finally, how "feelings" can be identified in this state structure. The latter calls on the concept of "Cognitive Phenomenology," which encompasses internal states that are phenomenological in a way that enhances classical notions of sensory phenomenology. The formal artefact used in the paper is the "neural automaton"[3] (or "neural state machine"), an abstraction drawn from the engineering of dynamic informational systems. The paper itself draws attention to important analytical outlines in the discovery of subjective feelings in machines.

MACHINE PHENOMENOLOGY

M is partitioned into those systems that have inner states, $M(I)$, (pendulums, state machines, brains . . . i.e., systems whose action is dependent on inner states that mediate perceptual input to achieve action) as against those that do not, $M(\sim I)$, (doorbells, perforated tape readers, translation machines . . . i.e., systems whose action depends on current input only). The "human machine" must belong to $M(I)$ and some of its inner states are the "mental" states that feature in definitions of human consciousness.

So, $M(F) \subset M(I)$ and to head towards a definition of $M(F)$, $M(I)$ needs refining, which comes from the fact that the inner state must be a subset of a phenomenological set of machines $M(P)$, that is, a set of machines in which the inner states can be *about* events in the world and for which *there is something describable it is like to be in that state*. That is, $M(F) \subset M(P)$, $M(P) \subset M(I)$.

Crucially, an "aboutness" in $M(P)$-type machines can be characterized as follows.

A particular machine A, where $A \in M(P)$, is influenced by a world, which in a simplified way, but one that does not distort the flow of the argument, produces a sequence of perceptual inputs to A

$$I^A = \{i_1^A, i_2^A, i_3^A, \dots\}$$

To be phenomenological, there needs to be a sequence of internal states in *A*

$$S^A = \{s_1^A, s_2^A, s_3^A, \dots\}$$

where s_j^A *is about* the corresponding i_j^A. This implies a coding that uniquely represents i_j^A. Indeed, the coding can be the same for the two or so similar as not to lose the uniqueness. This relationship is made physically possible at least through the *learning* property found in a neural state machine (or neural automaton) as pursued below.

ACHIEVING PHENOMENOLOGY IN NEURAL AUTOMATA

Here one recalls that in conventional automata theory the finite state dynamics of a general system from *M*(*I*) with inner states $\{a_1, a_2 \dots\}$ is described by the dynamic equation

$$a(t) = f[a(t-1), e(t)]$$

where *x*(*t*) refers to the value of a parameter at time *t* and *e*(*t*) is an external influence on the system, at time *t*. To aid the discussions and without the loss of relevance, time is assumed to be discretised. An *automaton* in condition [*a*(*t* – 1),*e*(*t*)] ***learns by internalizing*** *a*(*t*) to become an element of *f* in the sense that it "stores" *a*(*t*) as indexed by [*a*(*t* – 1),*e*(*t*)]. That is, given the automaton in [*a*(*t* – 1),*e*(*t*)], the next state entered by the automaton is the internalized state *a*(*t*). This storing function is achieved in a class of neural networks dubbed *neural automata* that are trained in a so-called *iconic* way.[4] The need to be neural has been central to this work as it provides generalization to cover cases similar but not identical to those encountered during learning.

Reverting to automaton *A*, say it is in some state s_{j-1}^A and receives the perceptual input i_k^A, then the dynamic equation may be rewritten to drop the superscript *A* as only one automaton is considered:

$$s_j = f[s_{j-1}, i_j]$$

To be *phenomenological* there needs to be a similarity relationship between s_j and i_j so that s_j can be said to *be about* i_j.

That is, using ≈ to mean "is equal to or uniquely similar to," then $sj \approx i_j$.[5]

This achieves a phenomenological relationship between *S* and *I*. Finally, it is noted that *f* is a mapping $S \times I \xrightarrow{f} S'$, where S′ is the set of "next" states while *S* is the set of current states.

ACHIEVING SUBJECTIVITY IN NEURAL AUTOMATA

So far, the automaton described is phenomenological to the extent that it has inner states that are about previously experienced external states. However, subjectivity (irrespectively of some differing definitions of what it means) includes the ability to make functional use of the created states "owned" by the entity in what would colloquially be called "thought." This first requires that internal phenomenological states can exist without the presence of input: a "perceptually unattended" situation. This input is given the symbol φ and is characterized by not creating a phenomenological state. So, say that the input i_a occurs more than once during the learning process then, starting in some state s_x when i_a occurs for the first time, we have

$$[s_x, i_a] \rightarrow s_a,$$

where (→) reads, "causes a transition to."

Then if φ is applied to the input, we have

$$[s_a, \varphi] \rightarrow s_a$$

The result of this entire action may be depicted by the commonly used state transition diagram (Figure 1.)

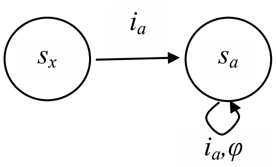


Figure 1: State diagram for the formation of subjective state s_a.

So with input φ, s_a becomes a self-sustained state which is *about* the last-seen input i_a. A further step is that the same φ can occur in the creation of any single phenomenological state so that the automaton may be said to *own* the inner version of all externally experienced single events.

But generally, experience consists of sequences of external influences and the current formulation needs to be extended to internal representations of time-dependent external experiences.

STATE STRUCTURES AND THOUGHT

To make experiences incurred in time subjective, consider the input changing from i_a to i_b. The relevant transition diagram then becomes (Figure 2).

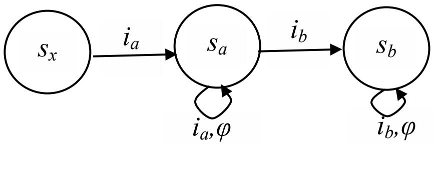


Figure 2: State diagram for the formation of the subjective experience of i_a followed by i_b.

To take this further, it is recalled that these behaviors are subjective to the extent that they "belong" to the automaton which physically performs the function

$$S \times I \xrightarrow{f} S'$$

where *f* is built up from experienced states and state transitions. It should be noted first that the automaton could be in some state s_p which on some occasions receives input i_q leading to s_q and other occasions i_r leading to s_r. Secondly, it is asserted (but can be shown to be true in specific cases) that the neutrality of φ is such that it allows transitions to each of the learned states in a probabilistic function. So, in the above examples, with φ as input, the automaton can change from state s_p to itself, s_q or s_r with probabilities determined by technological and learning exposure detail. The upshot of this is that the automaton develops a probabilistic structure of phenomenological states and transitions between them that are about past experience. This leads to the "ownership" of *explorable* internal state structure, which, in the case of living entities, is called **thought**. One's life, and that of an artificial entity, is based on a mix of inputs imposed by the world and φ, which allows thought to be driven by external perceptions, or previous internal states, that is, previous experience.

ATTRACTORS

Without going into detail about the statistical properties of neural networks, we note that for a particular input such as i_a in figure 2, there is only one state that remains sustained in time, and that is s_a. It turns out that for some neural networks (including natural ones) starting in a state that is not about , the states change getting more and more similar to s_a until s_a is reached. Then s_a is called an *attractor* under the input i_a. This issue returns in the consideration of volition below.

ACTION

The stated purpose of having subjective mental states is to enable the organism to act in some appropriate way in its world. This is closely connected to the concept of volition, as will be seen. Automata *action* is a concern in automata theory as, in general, an automaton, in addition to performing the next-state function $S \times I \xrightarrow{f} S'$ also performs an output function $S \xrightarrow{g} Z$ where Z is a set of output actions which in the most primitive entities, causes locomotion in its world. In more sophisticated entities language falls within the definition of Z. As with f, an automaton can learn to build up g as experience progresses. Here is an example. Say that the automaton can take four actions: movement in four cardinal directions, that is $Z = \{n,s,e,w\}$. The automaton can then either be driven in its (2D) world or it can explore it at random. In either case an element of $Z = \{n,s,e,w\}$ is associated with the state of the automaton and this determines the next input and associated state. Therefore, the state trajectory is now about a real world trajectory. The same principle applies to any other form of action, including language, in the sense that action, movement, utterances, or, indeed, inaction become associated with the state structure of the automaton leading, through the exploration of state trajectories to the ability to fix the brakes on a car, play the piano, or plan an escape from jail.

VOLITION AND ATTRACTORS

Referring to the paragraphs on attractors, the input or an internal state could represent something that is wanted. The resulting trajectory to an attractor in a system that performs actions internally represents the necessary actions for achieving the desired event. In the case of the automaton in the last section, this trajectory indicates the steps necessary to find that which is wanted. This is a substantial topic and previous literature on this functioning may be found.[6] In fact, this activity is part of a set of five requirements for the presence of consciousness in an automaton.[7] (Detail of this is not necessary for the current discussion.)

FEELINGS AND COGNITIVE PHENOMENOLOGY

It is the contention of a group of philosophers, Tim Bayne,[8] Galen Strawson,[9] and Michelle Montague,[10] that classical phenomenology is too closely allied to perceptual and sensory events and therefore avoids the discussion of mental states related to meaning, understanding, and abstract thought. Such states, it is argued, are *felt* alongside the sensory/perceptual. For example, were someone to utter a word in their own language, there is something it is like to understand such words; hence there is a cognitive character to this phenomenology. Advocates of cognitive phenomenology argue that this feeling is common to all utterances that are understood. Similarly, an utterance that is not understood is accompanied by a feeling that is common to all non-understood utterances Within our work with automata it has been suggested that feelings of understanding or not, the presence or absence of meaning in perceptual input, language understanding, and abstract thought are parts of the *shape* of state trajectories which affect the "what it's like" to be in these trajectories.[11] For example, a heard word that is understood will have a state trajectory that ends stably in an attractor. If not understood, the trajectory will be a random walk. In a machine, these differences in state behavior warrant different descriptions, which can be expressed in the action of the machine. This mirrors the way that perceptions and feelings warrant different actions in ourselves. Indeed, the effect of the two felt events on action can be very similar in machine and human. For example, an understood utterance (attractor) can lead to action whereas a non-understood one (random walk) may not.

SUMMARY AND CONCLUSION: A SEMANTIC EQUIVALENCE?

To summarize, in response to Koch's suggestion that there are no barriers to the human ability to engineer artefacts that have "subjective feelings," this paper has developed a *general* theoretical formulation that includes such systems. The salient points for this assertion are summarized below, followed by a discussion of arising issues.

I. The structure of the formulation is that of a formal, neural finite state machine (or automaton) which has perceptual input, inner states, and action dependent on the product of these.

II. The states of the machine are phenomenological by the system's *iconic* learning properties, making the internal states representative of experienced perceptual states.

III. Just having such coherent perceptual internal states is not sufficient to define mentation. The missing part is a linking *structure* of such phenomenological states that is also iconically learned from the actual history of experience of the automaton.

IV. *Actions* either taught or learned through exploration are associated with the states of the developing state structure.

V. A feature of the state structure is that it can be accessed internally within the artefact, in order to determine behavior in the world in which it is immersed, or deliberation without active behavior.

VI. To avoid the important observation that this approach does not cover phenomenal experiences such as "understanding" and "abstract deliberation" current work in progress on "cognitive phenomenology" has been outlined and argued to emerge from the *shape* of the state trajectories of the organism.

However, the above can be read as an engineering specification of an artefact that can operate in a progressively competent way in a world perceived by itself based on its learned dynamic phenomenological state structure **leaving the question of "subjective feelings" undebated**. It is the author's contention that the problem is a semantic one. In artificial systems the driving "mind," from the above, can be summarized as

a) "having an internally accessible structure of coherent stable perceptually experienced neural states."

It is posited that in a human being the descriptor "having subjective feelings" can be semantically expanded as follows. "Subjective feelings" in "qualia" discussions[12] are described as

b) "introspectively accessible, phenomenal aspects of our mental lives."

Here it is asserted that, semantically, the two descriptors are not meaningfully separable. The similarity between the artificial and the natural leads to the conclusion that the automata theory sequence in this paper describes artificial systems with "subjective feelings."

The above still leaves two issues that need to be clarified. The first is the possible casting of the human *qua* an "automaton," that is, an object without a conscious mind. This too is a semantic problem. The theory used in the paper is a very general way of describing dynamic information-based systems that, at a 1956 conference, was given the name "Automata Studies."[13] Such studies were intended to include methods of providing mathematical descriptions of human informational activities without treating the human in a diminished way.

The second is that it needs to be stressed that the material presented in this paper is **not about** an automaton that is **designed to be like a human being**. It is about a theory that is intended to represent formally and abstractly the "mental" activity of any organism thought to have a mental life. Drawn from the engineering of informational systems, the theory is not only intended to find the similarities between artificial and natural systems, but also draw attention to the differences. This makes possible the engineering of systems with usable mental lives, while also providing insights into how we might have a theoretical discussion about our personal subjective feelings.

Within the conceptual framework of engineering machine consciousness, engineering theories that are employed with informational systems provide a grounded language with which to discuss seemingly hard issues. In particular, it is stressed that informational theories are closer to the provision of the grounding needed than that provided by the classical physical sciences.

NOTES

[1] Igor Aleksander, *Impossible Minds: My Neurons, My Consciousness, Revised Edition* (London: Imperial College Press, 2015), 10.

[2] Stan Franklin et al., "A Phenomenally Conscious Robot?"

[3] Aleksander, *Impossible Minds*, 97.

[4] Ibid., 151.

[5] Technologically, this can easily be achieved by causing i_k^A to be a pattern of discrete values on a register to the terminals that represent the state of the system. Then similarity can be defined through the difference between such patterns on a point-to-point basis.

[6] Aleksander, *Impossible Minds*, 181–89; Aleksander, *The World In My Mind, My Mind in the World*, 130–39.

[7] Ibid., 29–39.

[8] Tim Bayne and Michelle Montague, *Cognitive Phenomenology*, 1–35.

[9] Ibid., 285-325.

[10] Michelle Montague, "Perception and Cognitive Phenomenology," 2045–62.

[11] Aleksander, "Cognitive Phenomenology: A Challenge for Neuromodelling," 395–98.

[12] Michael Tye, "Qualia," introductory paragraph.

[13] Claude Shannon and John McCarthy, *Automata Studies*.

BIBLIOGRAPHY

Aleksander, Igor. "Cognitive Phenomenology: A Challenge for Neuromodelling." *Proc. AISB*, 2017.

———. *Impossible Minds: My Neurons My Consciousness, Revised Edition*. London: Imperial College Press, 2015.

———. *The World In My Mind, My Mind in the World*. Exeter: Imprint Academic, 2005.

Bayne, Tim, and Michelle Montague. *Cognitive Phenomenology*. Oxford: Oxford University Press, 2011.

Franklin, Stan, Bernard Baars, and Uma Ramamurthy. "A Phenomenally Conscious Robot?" APA Newsletter on Philosophy and Computers 08, no. 1 (Fall 2008): 2–4.

Montague, Michelle. "Perception and Cognitive Phenomenology." *Philosophical Studies* 174, no. 8 (2017): 2045–62.

Shannon, Claude, and John McCarthy. *Automata Studies*. Princeton, Princeton University Press, 1956.

Tye, Michael. "Qualia." *The Stanford Encyclopedia of Philosophy*. Summer 2018 Edition.

Conscious Machine Perception

Magnus Johnsson

MALMÖ UNIVERSITY, SWEDEN; MOSCOW ENGINEERING PHYSICS INSTITUTE, RUSSIA; MAGNUS JOHNSSON AI RESEARCH AB, HÖÖR, SWEDEN

INTRODUCTION

An artificial cognitive architecture could be built by modeling the mammal brain at a systems level. This means that, though not modeling crucial components and their interconnections in detail, general principles also adhered to by their biological counterparts should be identified and followed in the design of such a system.

Systems-level modeling means identifying the components of the brain and their interactions. The components' functionality can then be implemented with mechanisms that model the systems at a suitable level of accuracy. The components can be re-implemented by other mechanisms for accuracy and performance reasons, or if more efficient implementations are found.

We could go about working on a bio-inspired systems-level cognitive architecture in various ways. At one extreme, we could work from a more holistic starting point by identifying crucial components and interactions found in the neural systems of biological organisms. Then we could implement maximally simplified versions of these and try to make them work together as well as possible. Examples of components in such an architecture inspired by a mammal brain could be a maximally simplified visual system and a maximally simplified mechanism, or set of mechanisms, corresponding to the Basal ganglia, etc. Inspired by the work of Valentino Braitenberg and the robotics physicist Mark W. Tilden, I believe such simplified but complete cognitive architectures would still enact interesting behaviors.

At the other extreme, we could work on individual components while trying to optimize these to perform at a human level or beyond. Many artificial perception researchers work at this extreme, e.g., by creating computer vision systems that in some respects even exceed the abilities of humans.

My approach is somewhere in the middle. I try to figure out general principles for not necessarily complete, but more composed architectures at an intermediary level. Hence my focus is not whether component implementations are optimized for performance. Following a systems-level approach, individual components can be reimplemented iteratively at later stages for performance, accuracy, or for other reasons, but this is not the focus here. Thus, the work on general principles can be isolated from the engineering questions of performance.

The perceptual parts of a cognitive architecture built according to these ideas employ to a large extent self-organizing topographical feature representations. Such feature representations are somewhat reminiscent of what has been found in mammal brains. These topographical representations are connected hierarchically, laterally, and recurrently. Hierarchies of increasingly complex feature representations—and in the extension different architectural components, possibly distributed—self-organize while supervising each other's associative learning/adaptation over space (by lateral associative connections) and over time (by recurrent associative connections). For example, such an architecture contains hierarchies of topographically ordered feature representations within sensory submodalities. To an extent, these hierarchies also cross the borders of different sensory submodalities, and even the borders of different sensory modalities. The topographically ordered feature representations connect laterally at various levels within, but also across, sensory submodalities, modalities, and to systems outside the perceptual parts, e.g., motor representations.

Below I discuss some principles that I believe are substantial to perception, various kinds of memory, expectations and imagery in the mammal brain, and for the design of a bio-inspired artificial cognitive architecture. I also suggest why these principles could explain our ability to represent novel concepts and imagine non-existing and perhaps impossible objects, while there are still limits to what we can imagine and think about. I will also present some ideas regarding how these principles could be relevant for an autonomous agent to become p-conscious[1] in the sense defined by Bołtuć,[2] i.e., as referring to first-person functional awareness of phenomenal information. Whether such an autonomous agent would also be conscious in a non-functional first-person phenomenological sense, i.e., h-conscious, adopting again the terminology of Bołtuć, and thus experience qualia of its own subjective first-person experiences of external objects and inner states, is another matter. The latter question belongs to the hard problem of consciousness.[3] The difficulty with that problem is that a physical explanation in terms of brain processes is an explanation in terms of structure and function, which can explain how a system's behavior is produced, but it is harder to see why the brain processes are accompanied by subjective awareness of qualia. According to Chalmers all metaphysical views on phenomenal consciousness are either reductive or nonreductive, and he considers the latter to be more promising. Nonreductive views require a re-conception of physical ontology. I suggest that the bio-inspired principles proposed in this paper have relevance for p-consciousness. Hence a cognitive architecture employing these ideas would probably become at least p-conscious. However, it is possible that h-consciousness is not a computational process, and I will not take a final position on the issue of phenomenal h-consciousness in this paper.

TOPOGRAPHICALLY ORDERED FEATURE REPRESENTATIONS

Topographically ordered maps are inherent parts of the human brain. There are continuously ordered representations of receptive surfaces across various sensory modalities, e.g., in the somatosensory and visual[4] areas, in neuron nuclei, and in the cerebellum.

The size of the representational area in such ordered representations depends on the behavioral importance and frequency of the represented input. For example, the representation of the fovea is much larger than the rest of the retina, and the representation of the fingertip is proportionally larger than the rest of the finger.

There are also more abstract topographically ordered representations in the brain, e.g., frequency preserving tonotopic maps[5] in primary auditory areas, and color maps in V4[6] in the visual areas.

In a model of such self-organized topographically ordered representations, essential relations among data should be made explicit. This could be achieved by forming spatial maps at an appropriate abstraction level depending on the purpose of the model. For a reasonable computational efficiency, the focus should be on main properties without any accurate replication of details. Reasonable candidates for a basic model corresponding to a topographically ordered representation in the brain satisfying these conditions are the Self-Organizing Map, SOM,[7] and its variants. Such a basic model forms a fundamental building block—not to be confused with the crucial components discussed above—in the perceptual parts of a bio-inspired cognitive architecture.

Examples of suitable candidates, beside the SOM, are the Growing Grid[8] and the Growing Cell Structure.[9] In addition to the adaptation of the neurons, these models also find suitable network structures and topologies through self-organizing processes. Other examples are the Tensor-Multiple Peak SOM, T-MPSOM,[10] or the Associative Self-Organizing Map.[11] The latter, or rather the principles it instantiates, are crucial for the principles of the perceptual parts of a cognitive architecture discussed in this paper and will be elaborated on below.

The SOM develops a representation that reflects the distance relations of the input, which is characteristic of lower levels of perception. If trained with a representative set of input, the SOM self-organizes into a dimensionality reduced and discretized topographically ordered feature representation also mirroring the probability distribution of received input. The latter means that frequent types of input will be represented with better resolution in the SOM. This corresponds to, for example, the development of a larger representational area of the fingertip than the rest of the finger in the brain, which was discussed above. Hence the SOM is reminiscent of the topographically ordered representations found in mammalian brains.

In a sense, the topographically ordered map generated by a SOM—and in the extension an interconnected system of SOMs—is a conceptual space[12] generated from the training data through a self-organizing process.

Due to the topology-preserving property of the SOM similar input elicit similar activity, which provides systems based on the SOM with an ability to generalize to novel input.

A SOM can be trained to represent various kinds of features, including phenomenal ones. The latter would turn the SOM into a phenomenal content map.[13] For example, a SOM can be trained to represent directions of lines/contours (as in V1), colors (as in V4), or more complex features such as the postures and gesture movements of an observed agent,[14] or the words of a text corpus ordered in a way that reflects their semantic relations.[15] Employing SOMs or other topographically ordered feature representations to represent phenomenal features together with the general design principles for a bio-inspired cognitive architecture suggested in this paper, would enable strong semantic computing.[16]

Other models of a self-organizing topology preserving feature representation are possible and might turn out to be more suitable for various reasons such as performance and accuracy. However, as also mentioned above, that is beyond the point of this paper, which aims at presenting higher level architectural principles where models of self-organizing topographically ordered representations are building blocks. Since I adhere to a systems-level modeling approach, subsystems of the cognitive architecture can be updated and substituted in an iterative fashion for improvement.

HIERARCHICAL FEATURE REPRESENTATIONS

How can self-organized topographically ordered representations of a more abstract kind, e.g., a representation with semantically related symbols that occupy neighboring places, be obtained in a cognitive architecture?

In the mammal brain there seems to be a principle of hierarchical ordering of representations, e.g., the executive and motor areas seem to be hierarchically ordered from more abstract to less abstract representations. Constraining the discussion to the perceptual parts of the mammal brain, we find that the different sensory modalities (visual, somatosensory, auditory, . . .) adhere to a hierarchical organizational principle. For example, we find hierarchically organized topology and probability-density preserving feature maps in the ventral visual stream of the visual system. These feature maps rely on the consecutive input from each other and tend to be hierarchically ordered from representations of features of a lower complexity to representations of features of a higher complexity. Thus, we find ordered representations of contour directions in V1 in the occipital lobe, of shapes in V2, of objects in V4, and of faces or complex facial features in the inferior temporal (IT) area of the temporal lobe.

The hierarchical organization principle is employed artificially in Deep Neural Networks, i.e., in artificial neural networks with several hidden layers. A neural network that has been applied very successfully within the field of computer vision is the Deep Convolutional Neural Network.[17]

Here, when I discuss the hierarchical ordering principle for perceptual parts of a bio-inspired cognitive architecture, this principle is instantiated by hierarchical SOMs. The choice of SOMs is not based on performance, but on the fact that the hierarchical organization principle is also to be combined with other principles in the cognitive architecture elaborated on below. For the moment the SOM and its variants are considered good choices to explain and test principles.

Together with collaborators, the author has shown the validity of this hierarchical organizational principle repeatedly with hierarchical SOMs when applied to different sensory modalities. For example, in the case of the somatosensory modality, several experiments have been conducted to show how haptic features of an increasing complexity can be extracted in hierarchical self-organizing representations, e.g., from proprioceptive and tactile representations at the lower complexity end to self-organizing representations of shapes and sizes of the haptically explored objects.[18] Another example in the case of the visual domain where experiments have been done to show that hierarchies of ordered representations of postures at the lower complexity end to ordered representations of gesture movements of the observed agent can be self-organized.[19]

LATERALLY AND RECURRENTLY CONNECTED FEATURE REPRESENTATIONS

The afferent signals from the sensory organs are a source of the perceptual activity in the brain. However, it is argued that a crucial aspect of biological cognition is an ability to simulate or influence perceptual activity in some brain areas due to the activity in other brain areas,[20] e.g., the activity in areas of other sensory modalities. For example, when the visual perception of a lightning evokes an expectation of the sound of thunder, or when visual images/expectations of an object is evoked when its texture is felt in the pocket. A more dramatic illustration is the well-known McGurk-MacDonald effect.[21] If a person sees a video with someone making the sound /da/ on which the lips cannot be seen closing and the actual sound played is /ba/, the expectations evoked by the visual perception may have such an influence on the activity caused by the actual afferent auditory sensor signals that the person may still hear the sound /da/.

Although there are hierarchically organized feature representations in the brain, it is questionable whether there are neurons—aka grandmother cells—that are the exclusive representatives of distinct individual objects. Though there is no total consensus regarding this, I consider it more likely that distinct individual objects are coded in a more distributed way as an ensemble of feature representations, at various complexity levels, across several sensory (as well as non-sensory) modalities. Hence, the recognition of distinct individual objects consists in the simultaneous activation of a sufficiently large and unique subset of this ensemble of representations across various modalities.

The activation of some feature representations will tend to trigger expectations/imaginations of features of the distinct individual object in other representations across various modalities, probably associated by lateral connections in a way similar to the activation of more features of higher—or lower—complexity in hierarchically connected feature representations (which can as well be cross-modal).

I believe that such associative connectivity, lateral and hierarchical, between feature representations of various modalities and at various complexity levels are what enables the filling in of missing parts of our perception by imagery, and that they enable our various kinds of memories.

Different kinds of memory are, I believe, using the same kind of feature representations across modalities in our brains. What differs between different kinds of memory is rather how they are activated and precisely what ensemble of feature representations that are activated. For example, one could speculate—in a simplified way—that the working memory consists in the activation of some ensemble of feature representations by executive circuits in the frontal lobes, whereas perception as such is the activation of an ensemble of feature representations due to afferent sensory signals, together with the filling-in of missing parts due to cross-modal as well as top-down expectations at various levels of hierarchies. Episodic memory, as well as imagery, could be the activation of the same/or similar ensembles as those activated by afferent sensory signals in the case of perception, but activated by neural activity within the brain. Semantic memory could be ensembles of feature representations that are subsets of those that make up episodic memories but with strengthened connections due to a reoccurring simultaneous activation during many various perceptual and episodic memory activations over time.

The point here is that all kinds of memory, perceptions, imaginations, and expectations are proposedly using simultaneous and/or sequential activations of ensembles/subsets of the same huge number of feature representations across various modalities in the brain. I think that there is no reason that the representations should be constrained to the brain only, but that associated representations could also be various kinds of activity in/of the body, e.g., postural/breathing patterns, hormonal configurations, etc. This would also explain why the change of posture/breathing patterns can change the state of the mind. In the extension, even "representations" that we interact with—and continuously reconfigure—in the environment outside the body—including the representations within other agent, such as humans, pets, machines, etc.—are probably included.

This kind of feature ensemble coding also enables/explains the ability to represent completely novel categories/concepts in the brain/cognitive architecture, and the ability to create and imagine non-existing and perhaps impossible concepts, objects, etc. This is because representations are composed of sufficiently large ensembles of associated multi-modal features, and novel associated ensembles and sometimes associated ensembles corresponding to concepts and imaginations that do exist (but have not yet been seen or reached) or do not exist in our physical reality (e.g., unicorns) can emerge.

Of course, there are limits to what we can imagine and conceptualize, and perhaps even think about. For example, we are unable to visualize objects in spaces of a higher dimensionality than three. However, such limitations are just to be expected if all perceptions, memories, and imaginations are made up of distributed (in space and time) activations of ensembles of associated features, and there are constraints on what kind of features can be represented in the brain (or cognitive architecture), which is likely. The constraints are probably set by biological limitations that exist due to a lack of evolutionary pressure, as well as determined by the development of the organism in its environment. An example of the latter is that cats raised in an environment consisting entirely of vertical lines during a critical developmental phase during infancy will be unable to see horizontal lines.[22] That there are constraints on what kind of features that can be represented also implies the possibility that all that we can think about regarding reality does not necessarily correspond to all that there would have been to think about had we been wired differently.

In accordance with the reasoning above, it is reasonable to assume that the need for lateral connections—corresponding to axon bundles in the neural system of a biological organism—between feature maps at various complexity levels within as well as between different modalities are of significance in a cognitive architecture based on self-organizing topographically ordered feature representations. Such lateral connections need to be adaptive (by adjustable parameters corresponding to modifiable synapses in the neural system of a biological organism) to enable the learning of associations between the activity in various feature representations.

In addition to an ability to automatically develop, and continuously readapt, sensory and other representations, and their interconnections that connect simultaneous activity within them spatially, a bio-inspired autonomous agent needs an ability to learn to associate activations of representations over time. This is desirable because it enables the autonomous agent to remember and re-enact sequences of perceptual—and other—activity across modalities and levels of hierarchy.

With such an ability an autonomous agent can remember sequences of perceptions, and if the ability is generalized, other things as well, e.g., motor activities. Such perceptual sequences could, for example, be visual landmarks. To the perceived visual landmarks, appropriate motor activity could be associated. With perceptual sequences simultaneously learned in other modalities together with cross-modal associations, the sequential memories are reinforced and thus diminish the influence of noise and limitations in sensory input. The perceptions (and preparatory responses, etc.) corresponding to missing sensory input in some modalities—sensory and other—will be imagined, i.e., elicited through cross-modal activation. If suddenly the agent would lack input to some, or all, sensory modalities, it would still be able to operate and to some extent carry out actions associated with imagined perceptions of the environment. With this kind of ability an agent would also be able to sit idle imagining various scenarios and the likely consequences of carrying out different kinds of actions. The latter is valuable for survival and will also accelerate the agent's learning.

The idea to internally elicit activity patterns in perceptual, motor, and other circuits over time (activation sequences) and in space (in various feature maps across different modalities), corresponding to the patterns that would have been elicited had there been sensory input and had the actions been carried out, is closely related to the simulation hypothesis by Hesslow.[23] It could in the extension also be the foundation for providing agents with an ability to guess the intentions of other agents, either by directly simulating the likely perceptual continuations of the perceived behavior of an observed agent, or by internally simulating its own likely behavior in the same situation under the assumption that the other agent is similar in its assessments, experiences, and values that drives it.

A mechanism that implements self-organizing topographically ordered feature representations that can be associatively connected with an arbitrary number of other representations, laterally and recurrently with arbitrary time delays, is the Associative Self-Organizing Map (A-SOM). Hence the A-SOM would in some cases be a better choice, than the standard SOM, to use as one of the basic building blocks in the perceptual parts of the cognitive architecture. An A-SOM can learn to associate the activity in its self-organized representation of input data with arbitrarily many sets of parallel inputs and with arbitrarily long-time delays. For example, it can learn to associate its activity with the activity of other self-organization maps, or with its own activity at one or more earlier times. This allows for cross-modal expectations. For example, if a sensory modality, say the visual system in a cognitive architecture, produces a certain internal pattern of activity due to sensory input, then activity patterns are elicited in other sensory modalities corresponding to the patterns of activity that are often triggered in these other sensory modalities through sensory inputs that usually occur simultaneously, even when they do not. Due to the ability of the A-SOM to associate its activity with its own activity at one or more earlier times, a mechanism for sequence completion that can be used for internal simulation is made possible. This is consistent with those abilities necessary for an autonomous agent described above. The A-SOM has been successfully tested in many simulations[24] in several different domains, as well as together with real sensors such as tactile sensors[25] and cameras,[26] and when simulating likely continuations of sequences of strings of symbols and words.[27] It has been used to simulate the sensory activity patterns likely to follow some initially perceived movements of actions/gestures.[28] In the domain of music, a further developed and more mature and generalized version of the A-SOM has been used to simulate the sensory activity patterns likely to follow those elicited by the initial parts of perceived Bach chorale melodies.[29]

Associative connections are in place between different representations at various levels of feature complexity. Simultaneously activated feature representations develop stronger associative connectivity. The result is that we will find strongly interconnected sets of feature representations—and other kinds of circuits—in the brain/architecture. As humans, we label these and call them systems/components of one kind or another, though we should keep in mind that these categorizations and demarcations are our inventions and thus somewhat arbitrary.

The inter-connectivity of the feature representations within a modality/submodality tend to be strong because it has been reinforced by simultaneous activations originating from the receptors of the modalityspecific sensory organs. Thus, connective configurations/subsystems in the brain/architecture develop through the repeated simultaneous activation of sets of self-organizing feature representations.

However, the feature representations within a modality also connect to feature representations in other modalities/systems, only to a lesser extent. This is due to the statistically fewer simultaneous activations of feature representations in other modalities. Various systems activate each other through these associative connections that have learned to associate activity that normally come together. Hence, if the activity within one system, perhaps triggered through afferent signals from sensory organs or from some other part of the brain/architecture, tend to correlate with the activity of other systems, perhaps triggered by the afferent signals from other sensory organs or other parts of the brain/architecture, then the inter-connectivity of the systems is reinforced. The foundation for these correlated activities in various systems is that sensory stimuli, and the consequences of an agent's actions, are related in a non-random way due to the statistical regularities of the properties of the world. These statistical regularities will be reflected in the associative connectivity between various systems.

Taken together, all this means that the activity in systems that are associatively connected to other systems in the brain/architecture also represent activity of—or what's going on—in the other systems.

7. CONSCIOUSNESS AND THE CREATION OF AN INNER WORLD

The perceptual parts of a cognitive architecture are those that are most relevant when it comes to consciousness. This is because consciousness is about something that is experienced. Hence, in the following, I will continue to constrain the discussion to perceptual parts.

In the discussion about cross-modal expectations and internal simulations above, I discussed how activity in some feature representations can elicit reasonable activity in other feature representations through associative connections. The elicited activity in the latter representations correspond to the activity that normally would or could occur simultaneously, or timed, with the activity in the first representations even though the latter lack any afferent input ultimately originating from sensors.

I believe that the same mechanism with adaptive associative connections in the case of a bio-inspired cognitive architecture, or nerve bundles with synapses in the case of a neural system of a biological organism, between different subsets of feature representations, at various levels of abstraction, is significant for the realization of at least p-consciousness. From this perspective, the elicitations of activity in some feature representations by the activity in other feature representations via associative connections can be viewed as if the activity in the latter system (composed of connected, perhaps distributed, feature representations) is observed by the former system. Various systems could perhaps also "observe" each other simultaneously as well. The mechanisms and principles sketched above could be used for a kind of summarization of the observed subsystem's or subsystems' activity at a possibly different and more abstract level.

As also argued by Hesslow and Jirenhed,[30] perceptual simulation could explain the appearance of an inner world. A remaining question is "who" is observing regardless of whether it is perceptions ultimately elicited from sensory organs, internal simulations originating from within the brain, or some combination thereof. My proposal is that they are observed by other connected configurations of systems whose activity summarizes/represents the observed internal simulations, because their corresponding activity correlates due to the learning represented in the adaptive associated connections. The same systems could perhaps have multiple functions while also "observing" each other simultaneously as well. Another way to put it is that some systems are aware of, i.e., p-conscious of, other systems' perceptual activity.

The activity of associatively connected configurations of feature representations correlate because the adaptations of the associative connections between the representations and the adaptions of the representations themselves happens simultaneously, continuously, and dynamically. At a lower perceptual level this means that the activation of feature representations in some sensory modalities will elicit activity in feature representations in other sensory modalities and consequently sensory expectations in those other modalities, as discussed above.

Thus, I believe that adaptive associative connections between and within various configurations of strongly connected feature representations at various levels of complexity or abstraction are of significant importance for realizing p-consciousness in a cognitive architecture.

NOTES

1. N. Block, "On a Confusion about a Function of Consciousness."
2. P. Bołtuć, "The Philosophical Issue in Machine Consciousness."
3. D. J. Chalmers, "Consciousness and Its Place in Nature."
4. D. van Essen, "Functional Organization of Primate Visual Cortex."
5. A. R. Tunturi, "Physiological Determination of the Arrangement of the Afferent Connections to the Middle Ectosylvian Auditory Area in the Dog," and "The Auditory Cortex of the Dog"; R. A. Reale and T. H. Imig, "Tonotopic Organization in Auditory Cortex of the Cat."
6. S. Zeki, "The Representation of Colours in the Cerebral Cortex."
7. T. Kohonen, *Self-Organization and Associative Memory*.
8. B. Fritzke, "Growing Grid — A Self-organizing Network with Constant Neighborhood Range and Adaptation Strength."
9. B. Fritzke, "Growing Cell Structures — A Self-organizing Network for Unsupervised and Supervised Learning."
10. M. Johnsson and C. Balkenius, "A Robot Hand with T-MPSOM Neural Networks in a Model of the Human Haptic System."
11. M. Johnsson et al., "Associative Self-organizing Map."
12. P. Gärdenfors, *Conceptual Spaces — The Geometry of Thought*.

13. A. Damasio, *Self Comes to Mind: Constructing the Conscious Brain*.
14. M. Buonamente et al., "Hierarchies of Self-organizing Maps for Action Recognition."
15. H. Ritter and T. Kohonen, "Self-organizing Semantic Maps."
16. P. Bołtuć, "Strong Semantic Computing."
17. Y. LeCun et al., "Gradient-Based Learning Applied to Document Recognition."
18. M. Johnsson et al., "Sense of Touch in Robots with Self-Organizing Maps."
19. M. Buonamente et al., "Hierarchies of Self-organizing Maps for Action Recognition."
20. G. Hesslow, "Conscious Thought as Simulation of Behaviour and Perception"; R. Grush, "The Emulation Theory of Representation: Motor Control, Imagery and Perception."
21. H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices."
22. C. Blakemore and G. F. Cooper, "Development of the Brain Depends on the Visual Environment."
23. G. Hesslow, "Conscious Thought as Simulation of Behaviour and Perception."
24. For example, M. Johnsson et al., "Associative Self-Organizing Map."
25. M. Johnsson and C. Balkenius, "Associating SOM Representations of Haptic Submodalities."
26. M. Buonamente et al., "Discriminating and Simulating Actions with the Associative Self-organizing Map."
27. D. Gil et al., "SARASOM – A Supervised Architecture Based on the Recurrent Associative SOM."
28. M. Buonamente et al., "Discriminating and Simulating Actions with the Associative Self-organizing Map."
29. M. Buonamente et al., "Simulating Music with Associative Self-organizing Maps."
30. G. Hesslow and D.-A. Jirenhed, "The Inner World of a Simple Robot."

BIBLIOGRAPHY

Blakemore, C., and G. F. Cooper. "Development of the Brain Depends on the Visual Environment." *Nature* 228 (1970): 477–78.

Block, N. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (1995): 227–87.

Bołtuć, P. "The Philosophical Issue in Machine Consciousness." *International Journal of Machine Consciousness* 1, no. 1 (2009): 155–76.

———. "Strong Semantic Computing." *Procedia Computer Science* 123 (2018): 98–103.

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

Buonamente, M., H. Dindo, and M. Johnsson. "Discriminating and Simulating Actions with the Associative Self-organizing Map." *Connection Science* 27, no. 2 (2015): 118–36.

Buonamente, M., H. Dindo, and M. Johnsson. "Hierarchies of Self-organizing Maps for Action Recognition." *Cognitive Systems Research* 39 (2016): 33–41.

Buonamente, M., H. Dindo, A. Chella, and M. Johnsson. "Simulating Music with Associative Self-organizing Maps." *Journal of Biologically Inspired Cognitive Architectures* 25 (2018): 135–40.

Chalmers, D. J. "Consciousness and Its Place in Nature." In *Blackwell Guide to the Philosophy of Mind*, edited by S. Stich and T. Warfield, 102–42. Malden, MA: Blackwell Publishing, 2003.

Damasio, A. *Self Comes to Mind: Constructing the Conscious Brain*. Pantheon, 2010.

Fritzke, B. "Growing Grid — A Self-organizing Network with Constant Neighborhood Range and Adaptation Strength." *Neural Processing Letters* 2, no. 5 (1995): 9–13.

———. "Growing Cell Structures — A Self-organizing Network for Unsupervised and Supervised Learning." *Neural Networks* 7, no. 9 (1994): 1441–60.

Gärdenfors, P. *Conceptual Spaces — The Geometry of Thought*. Cambridge, MA: The MIT Press, 2000.

Gil, D., J. Garcia, M. Cazorla, and M. Johnsson. "SARASOM – A Supervised Architecture Based on the Recurrent Associative SOM." *Neural Computing and Applications* 26, no. 5 (2014): 1103–15.

Grush, R. "The Emulation Theory of Representation: Motor Control, Imagery and Perception." *Behav. Brain Sci* 27 (2004): 377–442.

Hesslow, G. "Conscious Thought as Simulation of Behaviour and Perception." *Trends Cogn Sci* 6 (2002): 242–47.

Hesslow, G., and D.-A. Jirenhed. "The Inner World of a Simple Robot." *J. Consc. Stud.* 14 (2007): 85–96.

Johnsson, M., and C. Balkenius. "A Robot Hand with T-MPSOM Neural Networks in a Model of the Human Haptic System." In *The Proceedings of Towards Autonomous Robotic Systems* (2006): 80–87.

Johnsson, M., and C. Balkenius. "Associating SOM Representations of Haptic Submodalities." In *the Proceedings of Towards Autonomous Robotic Systems 2008* (2008): 124–29.

Johnsson, M., C. Balkenius, and G. Hesslow. "Associative Self-organizing Map." In *The Proceedings of the International Joint Conference on Computational Intelligence (IJCCI) 2009* (2009): 363–70.

Johnsson, M., and C. Balkenius. "Sense of Touch in Robots with Self-Organizing Maps." *IEEE Transactions on Robotics* 27, no. 3 (2011a): 498–507.

Johnsson, M., M. Martinsson, D. Gil, and G. Hesslow. "Associative Self-Organizing Map." In *Self-Organizing Maps – Applications and Novel Algorithm Design* (2011b): 603–26.

Kohonen, T. *Self-Organization and Associative Memory*. Berlin Heidelberg: Springer-Verlag, 1988.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* (1998): 2278–324.

McGurk, H., and J. MacDonald. "Hearing Lips and Seeing Voices." *Nature* 264 (1976): 746–48.

Reale, R. A., and T. H. Imig. "Tonotopic Organization in Auditory Cortex of the Cat." *J Comp Neurol* 192 (1980): 265–91.

Ritter, H., and T. Kohonen. "Self-organizing Semantic Maps." *Biol. Cybern* 61 (1989): 241–54.

Tunturi, A. R. "Physiological Determination of the Arrangement of the Afferent Connections to the Middle Ectosylvian Auditory Area in the Dog." *Am J Physiol* 162 (1950): 489–502.

———. "The Auditory Cortex of the Dog." *Am J Physiol* 168 (1952): 712–17.

van Essen, D. "Functional Organization of Primate Visual Cortex." *Cerebral Cortex* 3 (1985): 259–329.

Zeki, S. "The Representation of Colours in the Cerebral Cortex." *Nature* 284 (1980): 412–18.

Transhumanism: The Best Minds of Our Generation are Needed for Shaping Our Future

Stefan Lorenz Sorgner
JOHN CABOT UNIVERSITY, ROME

The great plurality of emerging technologies permanently raises new moral and anthropological challenges. With each invention, new challenges come about, which can be interpreted in many different ways. Which interpretation is the most plausible one? How should we legally deal with these new challenges? What are the corresponding economic implications? Do techniques alter who we are, or do they merely serve as means for realizing specific goals? These are some of the tricky issues with which all of us are being currently confronted. No one knows what will actually happen in the future. Yet, we are in a position to make decisions. The moral and anthropological issues related to emerging technologies are not only being discussed by young geeks or elderly experts, but all of us are permanently being confronted by the implications of emerging technologies. When we go to the cinema, the movie "Transcence" shows us what can be expected when singularity occurs.[1] When we watch the series "Black Mirror," we are being confronted with a version of the social credit system that has been implemented in China. When we read the novel "Inferno" by Dan Brown, we are being shown a technical solution concerning overpopulation.

Emerging technologies not only alter the world we live in, but they also have the potential to modify human beings such that the possibility arises that human beings realize their own self-overcoming. How should this option be evaluated, if we assume that we have an unchanging human nature? The belief in an eternal human nature used to be dominant for 2,500 years. In the past 200 years, it has been challenged by evolutionary thinking. Darwin and Nietzsche have brought about a new understanding of who we as human beings are.[2] This topic, too, needs to be taken into consideration, as we have not yet managed to grasp all the implications of their paradigm-shifting reflections.[3]

In 2004, the magazine "Foreign Policy" asked the leading intellectual and political scientist Francis Fukuyama what he regards as the world's most dangerous idea. His answer was transhumanism. Transhumanism is a cultural movement that affirms the use of techniques to increase the likelihood that human beings manage to transcend the boundaries of their current existence.[4] We should take evolution into our own hands. The term transhumanism was coined bv the first director of UNESCO, Julian Huxley, in 1951. Currently, transhumanist thinking is being affirmed by many futurists and innovators in the Silicon Valley, who are directly responsible for shaping the world by means of innovations.

In my own writings, I present a Nietzschean transhumanism.[5] By presenting a short summary of some of the implications of this way of thinking, the relevance and scope of dealing with emerging technologies becomes evident. The following thoughts represent my outlook concerning the most promising techniques that are available for altering who we are as human beings.[6] These reflections reveal the wide range of social, economic, cultural, anthropological, and moral challenges that go along with emerging technologies. Perhaps they will also serve as impulses for further intellectual debates.

We are on the way towards the posthuman. Nietzsche spoke about the overhuman, which was his way of presenting the same insight. We are completely part of nature, differ only gradually from other living beings, and are threatened with extinction like all other living beings. There are only two possibilities: Either we constantly evolve to adapt to our environment, or soon we will no longer exist. *Homo sapiens sapiens* is the result of a development from *Homo habilis* to *Homo erectus* and *Homo sapiens*. It would be naïve to assume that *Homo sapiens sapiens* is the crowning glory of this evolutionary development. If we do not want to become extinct, we will have to evolve. Otherwise, the environmental conditions that are constantly changing themselves will become hostile to us and cause our extinction. If we are successful, *Homo sapiens sapiens* will still be there for a while. Finally, *Homo sapiens sapiens*, too, will be replaced by the posthuman.

Biology professor Julian Huxley assumed exactly this when he developed the principles of his transhumanism. In order to promote adaptation to the environment, and at the same time personal well-being, transhumanist thought not only advocates the use of the latest techniques for human development, but also attaches great importance to education. Only in this way can the probability of the next evolutionary step towards the posthuman be increased.

The developments described are constantly in progress. We have various technical possibilities for promoting human flourishing. The field of genetic engineering is particularly promising with regard to the potential of the further development of humans. Bioprinters, Crispr/CAS9, PID, and 23andme are the decisive buzzwords here. Genetic modification of one's own offspring determined by parents, is structurally analogous to traditional education on closer analysis and should therefore also be evaluated analogously from a moral point of view.

Education has always included genetic modification. The latest insights from epigenetics (i.e., the study of environmentally caused alterations of genes) underline this assessment.[7] Due to the developments of recent years, in particular with regard to the development of CRISPR/Cas9, a cheap, precise and reliable so-called "gene scissor," and Big Gene Data (i.e., the application of Big Data analyses to genes), this subject area has gained enormously in relevance. The potential for promoting the emergence of a new species by means of a multitude of genetic techniques can hardly be overestimated. We are already in a position today to make selections following previous pre-implantation diagnosis as part of artificial fertilization. The ethical, political, and legal framework is the reason why we are not yet doing what we are already technically capable of doing.[8] The other two decisive technical possibilities to support the autopoietic self-overcoming process are the promotion of human-machine interfaces and artificial intelligence. I consider the human-machine interfaces in particular to be of central importance for human development because smart cities also need upgraded people. If all areas of life are equipped with an RFID chip, i.e., with (active RFID chips) or without (passive RFID chips) antennae radio frequency identification chips, then this must also be done with us humans in order to be able to guarantee efficient interaction. Computers are getting smaller in rapid steps. Twenty-five years ago we had PCs. These are increasingly being replaced by the smartphone. The next step on which companies are working already is to integrate the computer into people. The monitor is then coupled directly to our optic nerves. We operate it by gesture control. The text input takes place directly through our thinking. The future of writing is thinking. The Internet of Things is thus supplemented by the Internet of Bodily Things. Sensors of the integrated computer will be located in different parts of our body in order to be able to check our bodily functions. Researchers at Tufts University have already developed a sensor that can be integrated into our teeth to monitor our food intake.[9] Using these sensors and the permanent monitoring of our body, we can detect diseases not only when they are far advanced, but possibly even before they have begun to develop. Predictive maintenance is the name given to this process in machines. Predictive maintenance will also be possible in humans with the evolution of the Internet of Bodily Things, a network of interacting chips located in the human body, which in turn will radically increase the human health span, the span of a healthy life. Expanding the human health span is a central goal of most transhumanists.[10] I consider these visions of genetic development and the upgraded human being to be probable and promising.

Unfortunately, transhumanism in public perception is often associated with another technique and a particular philosophy of mind actively represented in the media by Elon Musk and his friends: mind uploading.[11] This is the idea that our personality will be loaded into a computer and that the future of human existence will be a digital one. The so-called simulation argument,[12] which is often discussed by Musk in public, presupposes the possibility of mind uploading and presents reasons why its realization is obvious. A predominant interpretation of Moore's law suggests, for example, that the processor power of computers doubles every two years. Since the human personality is to be understood as software that runs on the hardware of the body, it is to be expected that in the coming decades the performance of processors will be so high that human personalities can exist as software on computers. In this way, human immortality can be realized.[13] This is a highly implausible thought, as we have no reason for claiming that life and, in particular, consciousness can exist in a silicon-based entity. Unfortunately, it is exactly this thought that is primarily identified with transhumanism in the public: Transhumanists want to become immortal by means of mind uploading. This is how transhumanism is usually presented in the media. But this is not a characterization that affects all varieties of transhumanism. Numerous transhumanists do indeed proceed from the plausibility of this idea, since they share the image of human beings just mentioned. However, transhumanism is not necessarily linked to this anthropology, which implies a functional theory of mind.[14] It is definitely not a way to become immortal. We cannot even conceptualize immortality in a meaningful manner. The expansion process of our universe might stop eventually so that no further movements will occur. Alternatively, the expansion process of the universe could turn into a contraction process that could take us towards a cosmological singularity. How should human beings be able to survive such a process? Yet, this is what ought to be possible, if human immortality were an option.

The essence of transhumanism lies exclusively in affirming the use of the latest techniques to promote the likelihood of a good life. Which techniques are relevant here and which concept of the good life is supposed to be considered, is intensively discussed among transhumanists. I myself cannot rule out the possibility of mind uploading, but for numerous reasons I do not associate it with too much hope in the near future. The decisive reason is that all forms of life known to us are based on a carbon base, whereby the quality of life goes hand in hand with the ability of self-movement. In trees, self-movement consists of independent growth. This does not mean that I exclude the possibility that life can also exist on a silicon basis, but we do not currently have any evidence of this. All living entities are carbonate-based and all conscious beings known to us are alive. If there was at least one living entity which is silicon based, we can further consider the option of mind uploading. For this reason I consider the assumption that in thirty years we will be able to load our personality onto a hard disk as highly implausible. What I mean by this in a nutshell is the following. In the middle ages, scholars discussed how many angels fit on the tip of a needle. Nowadays, we talk about the simulation argument. Both topics are fun. Both discourses make sense from the perspective of the specific cultural background. Yet, in both instances we avoid being concerned with the most pressing issues of our times.[15]

Much more promising than the transfer of the personality to a computer seems to me to be the integration of the computer into the body, because it is this line of development that we have been able to observe for decades. Computers are becoming smaller and smaller and more and more integrated into the body. Such an upgraded body can in turn be well monitored by computers, through which we gain numerous insights into physical and genetic processes. In this way, too, the possibilities of human self-design can be promoted. In particular, the already reliable, precise, and cost-effective genetic modification techniques, which I consider to be the most important scientific innovations of this decade, will enable us in the not too distant future to overcome numerous previous limits of our humanity. However, we will not achieve immortality in this way.

Still, we need to seriously consider the implications of emerging technologies. The common ancestors we have with great apes lived six million years ago. The public use of the internet has been realized less than forty years ago. Promising genome editing techniques have only been realized recently. Yet, both techniques have significantly affected our lives already. We need to think about the implications of these technologies, as they concern all aspects of our life world. No one knows exactly what the future will lead to. Now, we can think about it, make decisions, and act accordingly. The best minds of our generation are needed to reflect upon the impacts of emerging technologies, and for actively shaping our future.

NOTES

1. See R. Kurzweil and T. Grossman, *Transcend. Nine Steps to Living Well Forever*.
2. S. L. Sorgner, *Menschenwürde nach Nietzsche: Die Geschichte eines Begriffs*.
3. Reflections on the impact of emerging technologies are of central relevance for shaping the world we live in. We even develop capacities for actively altering who we will be as human beings. Given the enormous potential of emerging technologies, we need to discuss what we want, and which values, and norms are supposed to be the principles on which we base our actions (Sorgner, *Transhumanismus: 'Die gefährlichste Idee der Welt'!?*). To be able to comprehensively reflect upon these issues, we need an appropriate basis for developing them. This is where

the need to expand the current curriculum of the humanities comes in, as the humanities in their traditional form are not well equipped for dealing with the great variety of challenges which go along with the latest technical developments. All of these reflections demonstrate why a meta-disciplinary intellectual forum for dealing with emerging technologies is urgently needed. Together with James Hughes, Sangkyu Shin, and Penn State University Press, I established the "Journal of Posthuman Studies" in 2017. It is the world's first double-blind peer review journal explicitly dedicated to the posthuman. The posthuman turn is a paradigm-shifting event. We have not even realized all the implications that come along with our new non-dualist understanding of being human as well as with the practical implications of emerging technologies yet.

4. R. Ranisch and S. L. Sorgner, *Introducing Post- and Transhumanism." In Post- and Transhumanism. An Introduction*.
5. S. L. Sorgner, *Übermensch. Plädoyer für einen Nietzscheanischen Transhumanismus*.
6. A detailed outline of this approach can be found in my recently published monograph "Schöner neuer Mensch" (2018).
7. S. L. Sorgner, "The Future of Education: Genetic Enhancement and Metahumanities."
8. J. Savulescu, "Procreative Beneficence: Why We Should Select the Best Children"; J. Savulescu and G. Kahane, "The Moral Obligation to Create Children with the Best Chance of the Best Life"; S. L. Sorgner, "Is There a 'Moral Obligation to Create Children with the Best Chance of the Best Life'?"
9. "Scientists Develop Tiny Tooth-Mounted Sensors That Can Track What You Eat," *TuftsNow*, May 22, 2018. Available at https://now.tufts.edu/news-releases/scientists-develop-tiny-tooth-mounted-sensors-can-track-what-you-eat
10. A. de Grey and M. Rae, *Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Aging in Our Lifetime.*
11. R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*.
12. N. Bostrom, "Are You Living in a Computer Simulation?"
13. M. Rothblatt, *Virtually Human. The Promise—and the Peril—of Digital Immortality*.
14. M. More, "The Philosophy of Transhumanism."
15. S. L. Sorgner, *Schöner neuer Mensch*.

BIBLIOGRAPHY

Bostrom, N. "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53, no. 211 (2003): 243–55.

de Grey, A., and M. Rae. *Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Aging in Our Lifetime*. New York: St. Martin's Press, 2007.

Fukuyama, F. "The World's Most Dangerous Ideas: Transhumanism." *Foreign Policy* 144 (2004): 42–43.

Huxley, J. "Knowledge, Morality, and Destiny—The William Alanson White Memorial Lectures, third series." *Psychiatry* 14, no. 2 (1951): 127–51.

Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*. London: Penguin, 2006.

Kurzweil, R., and T. Grossman. *Transcend. Nine Steps to Living Well Forever*. New York: Rodale Books, 2011.

More, M. "The Philosophy of Transhumanism." In *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*. Edited by M. More, N. Vita-More, 3–17. Chichester: Wiley-Blackwell, 2013.

Ranisch, R., and S. L. Sorgner. *Introducing Post- and Transhumanism." In Post- and Transhumanism. An Introduction*. Edited by R. Ranisch and S. L. Sorgner, 7–29. Peter Lang, Frankfurt a. M.: Peter Lang, 2014.

Rothblatt, M. *Virtually Human. The Promise—and the Peril—of Digital Immortality*. London: Picador, 2015.

Savulescu, J. "Procreative Beneficence: Why We Should Select the Best Children." *Bioethics* 15, nos. 5-6 (2001): 413–26.

Savulescu, J., and G. Kahane. "The Moral Obligation to Create Children with the Best Chance of the Best Life." *Bioethics* 23, no. 5 (2009): 274–90.

Sorgner, S. L. "Is There a 'Moral Obligation to Create Children with the Best Chance of the Best Life'?" *Humana Mente: Journal of Philosophical Studies* 26 (2014): 199–212.

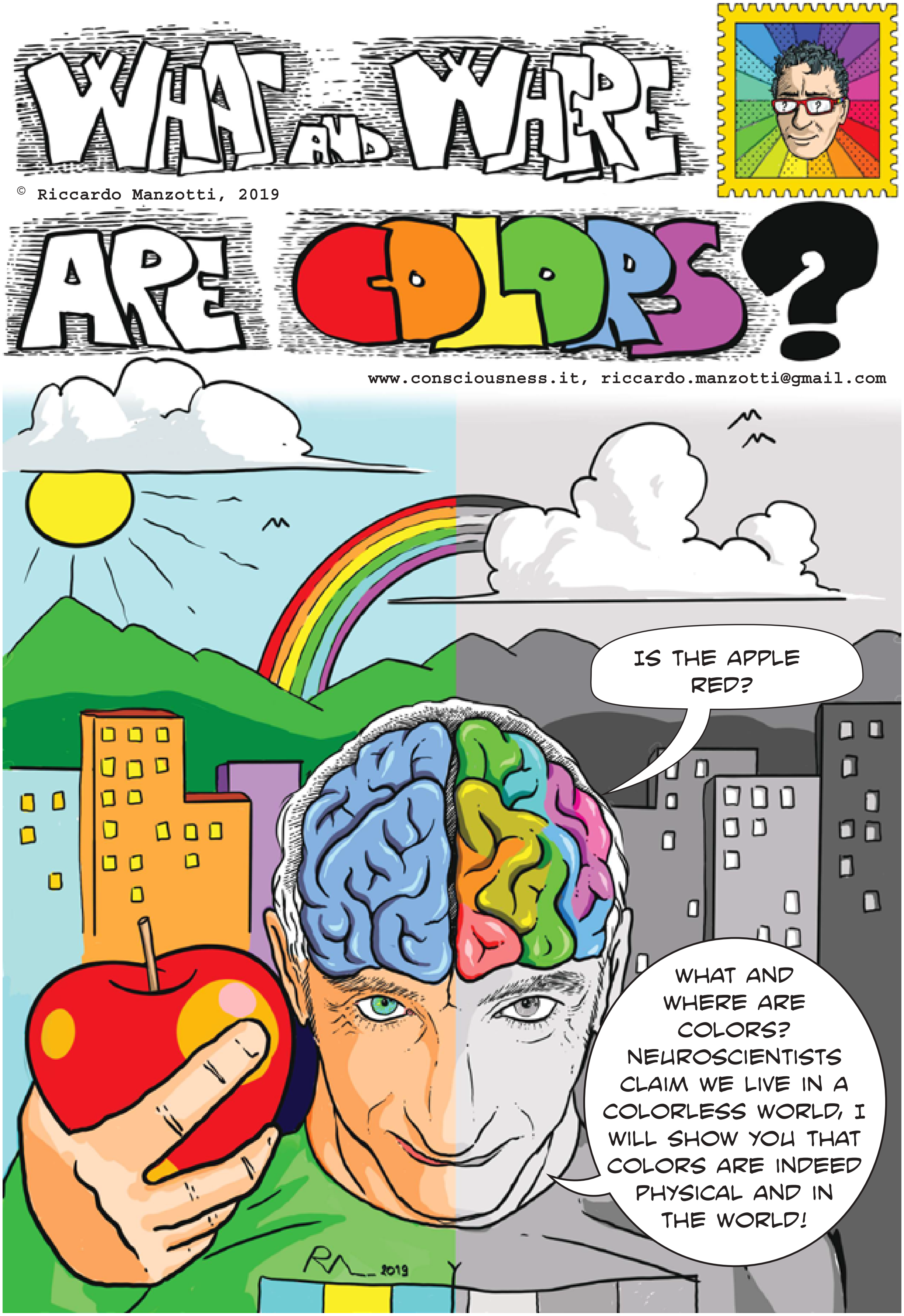
Sorgner, S. L. *Schöner neuer Mensch*. Berlin: Nicolai Verlag, 2018.

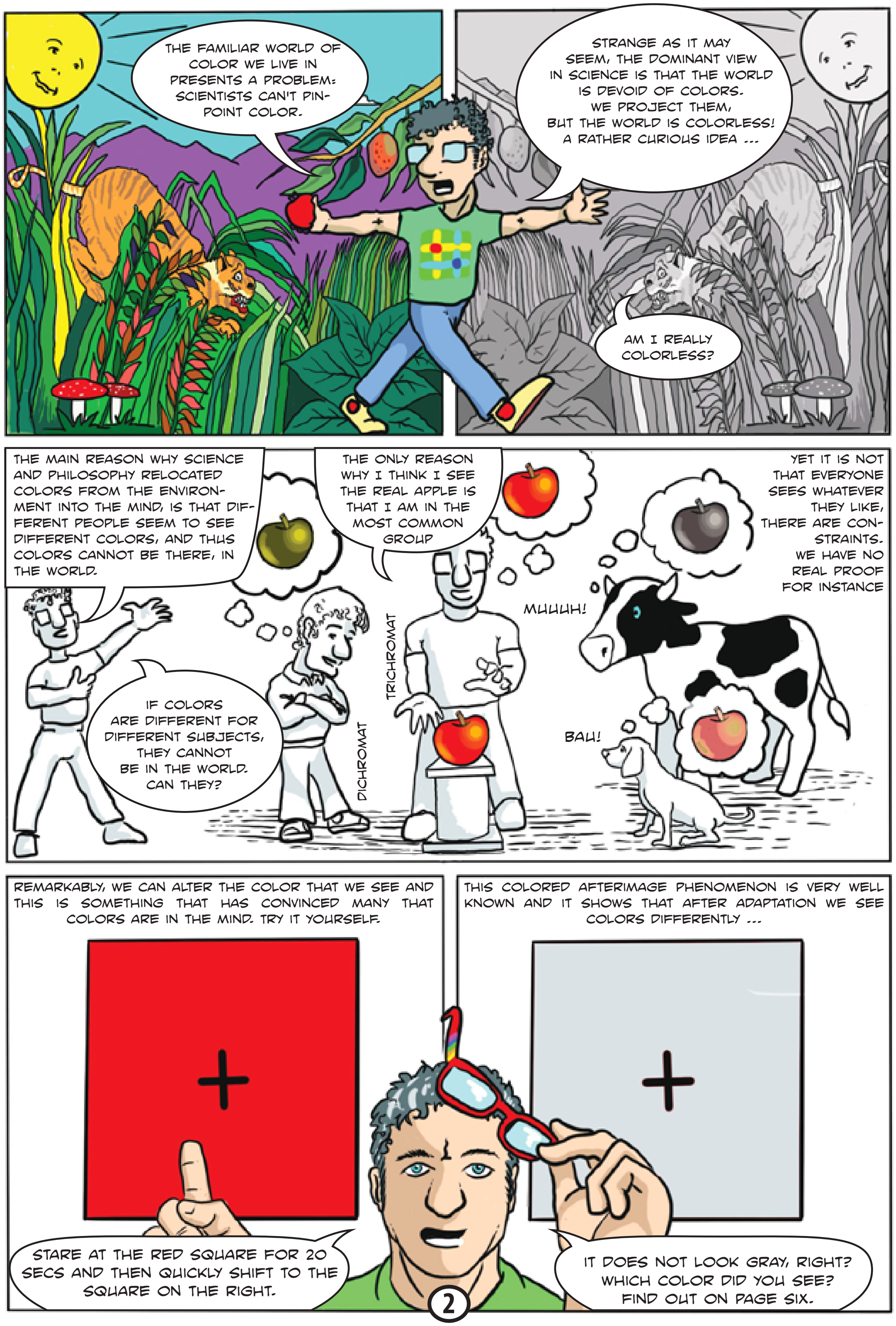
Sorgner, S. L. *Übermensch. Plädoyer für einen Nietzscheanischen Transhumanismus*. Basel: Schwabe, 2019.

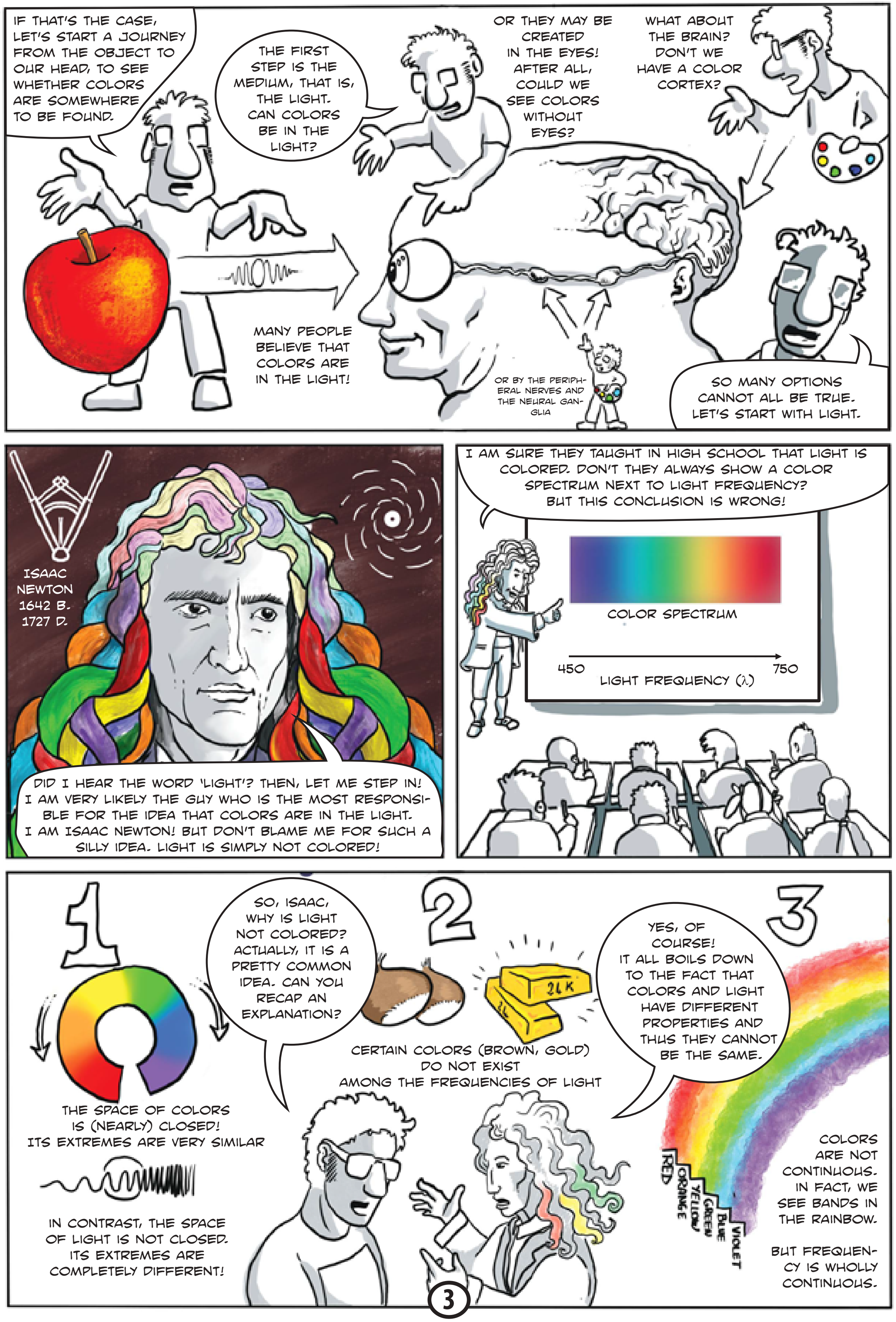
Sorgner, S. L. "The Future of Education: Genetic Enhancement and Metahumanities." *Journal of Evolution and Technology* 25, no. 1 (May 2015): 31–48.

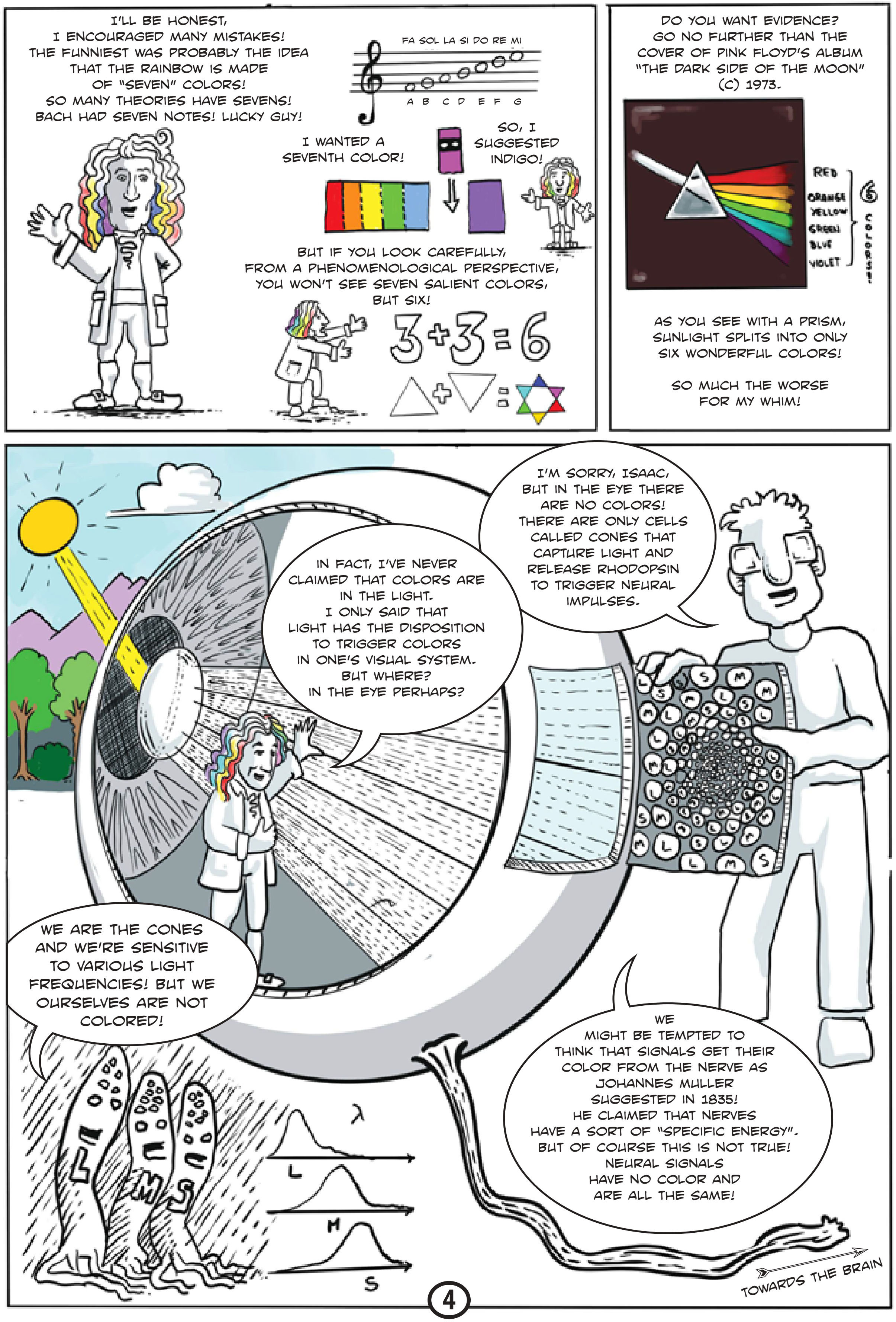
Sorgner, S. L. *Transhumanismus: 'Die gefährlichste Idee der Welt'!?* Herder, Freiburg i. Br.: Herder, 2016a.

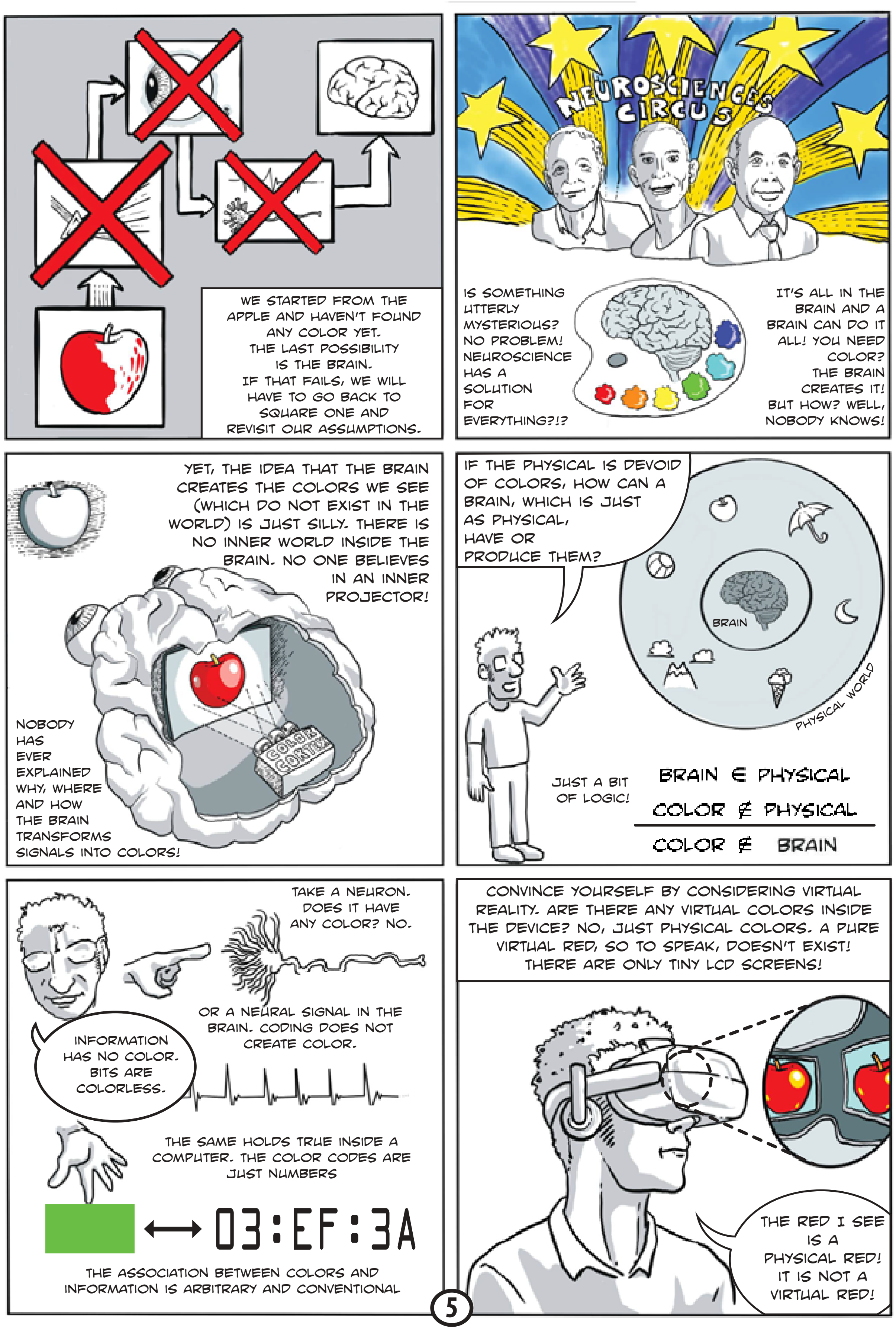
Sorgner, S. L. *Menschenwürde nach Nietzsche: Die Geschichte eines Begriffs*. WBG, Darmstadt: WBG, 2010.

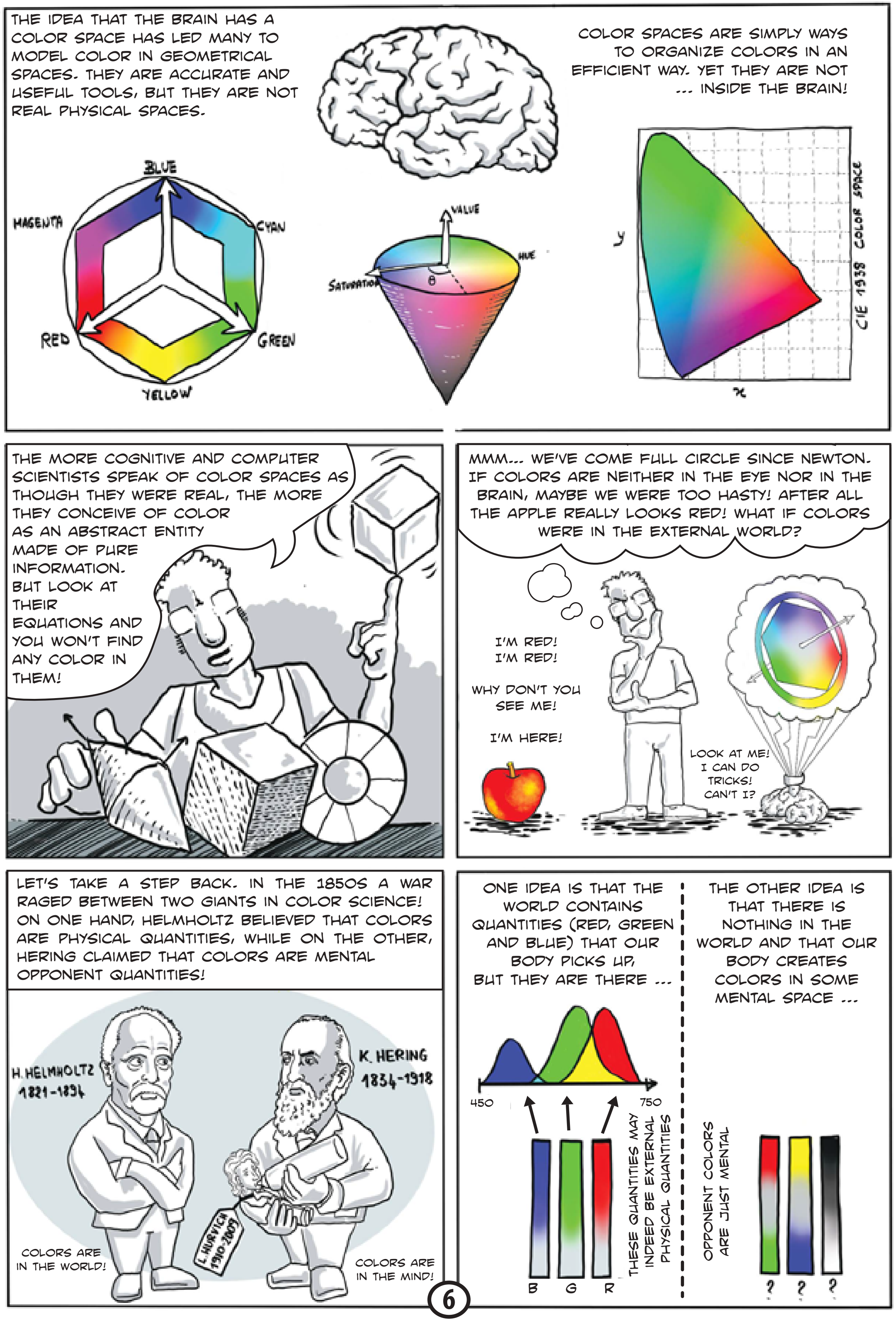

WHAT AND WHERE
© Riccardo Manzotti, 2019
ARE COLORS?
www.consciousness.it, riccardo.manzotti@gmail.com
IS THE APPLE RED?
WHAT AND WHERE ARE COLORS? NEUROSCIENTISTS CLAIM WE LIVE IN A COLORLESS WORLD, I WILL SHOW YOU THAT COLORS ARE INDEED PHYSICAL AND IN THE WORLD!


THE FAMILIAR WORLD OF COLOR WE LIVE IN PRESENTS A PROBLEM: SCIENTISTS CAN'T PIN-POINT COLOR.
STRANGE AS IT MAY SEEM, THE DOMINANT VIEW IN SCIENCE IS THAT THE WORLD IS DEVOID OF COLORS. WE PROJECT THEM, BUT THE WORLD IS COLORLESS! A RATHER CURIOUS IDEA ...
AM I REALLY COLORLESS?
THE MAIN REASON WHY SCIENCE AND PHILOSOPHY RELOCATED COLORS FROM THE ENVIRONMENT INTO THE MIND, IS THAT DIFFERENT PEOPLE SEEM TO SEE DIFFERENT COLORS, AND THUS COLORS CANNOT BE THERE, IN THE WORLD.
THE ONLY REASON WHY I THINK I SEE THE REAL APPLE IS THAT I AM IN THE MOST COMMON GROUP
YET IT IS NOT THAT EVERYONE SEES WHATEVER THEY LIKE, THERE ARE CONSTRAINTS. WE HAVE NO REAL PROOF FOR INSTANCE
MUUUH!
TRICHROMAT
IF COLORS ARE DIFFERENT FOR DIFFERENT SUBJECTS, THEY CANNOT BE IN THE WORLD. CAN THEY?
DICHROMAT
BAU!
REMARKABLY, WE CAN ALTER THE COLOR THAT WE SEE AND THIS IS SOMETHING THAT HAS CONVINCED MANY THAT COLORS ARE IN THE MIND. TRY IT YOURSELF.
THIS COLORED AFTERIMAGE PHENOMENON IS VERY WELL KNOWN AND IT SHOWS THAT AFTER ADAPTATION WE SEE COLORS DIFFERENTLY ...
STARE AT THE RED SQUARE FOR 20 SECS AND THEN QUICKLY SHIFT TO THE SQUARE ON THE RIGHT.
IT DOES NOT LOOK GRAY, RIGHT? WHICH COLOR DID YOU SEE? FIND OUT ON PAGE SIX.
2

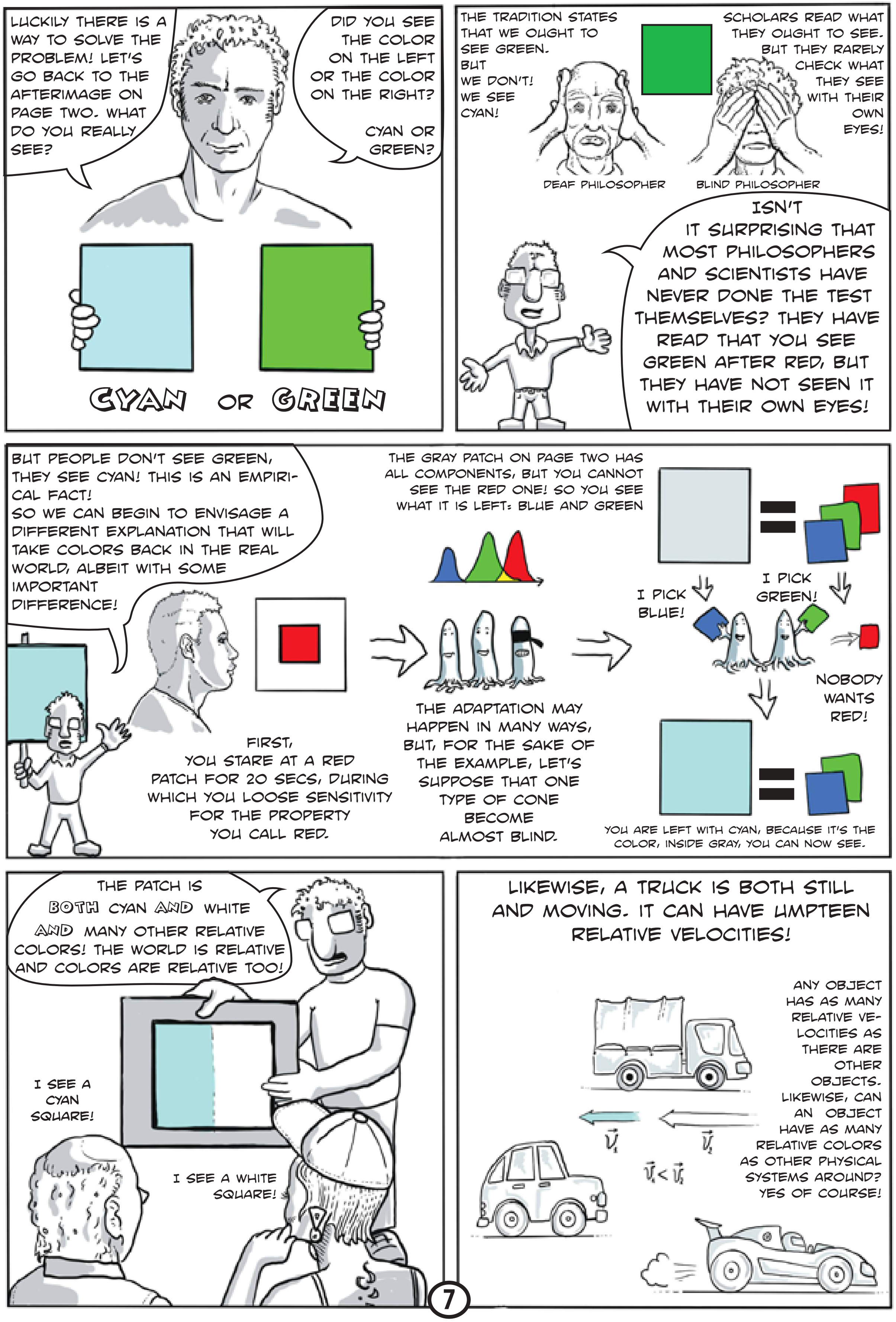

IF THAT'S THE CASE, LET'S START A JOURNEY FROM THE OBJECT TO OUR HEAD, TO SEE WHETHER COLORS ARE SOMEWHERE TO BE FOUND.
THE FIRST STEP IS THE MEDIUM, THAT IS, THE LIGHT. CAN COLORS BE IN THE LIGHT?
OR THEY MAY BE CREATED IN THE EYES! AFTER ALL, COULD WE SEE COLORS WITHOUT EYES?
WHAT ABOUT THE BRAIN? DON'T WE HAVE A COLOR CORTEX?
MANY PEOPLE BELIEVE THAT COLORS ARE IN THE LIGHT!
OR BY THE PERIPHERAL NERVES AND THE NEURAL GANGLIA
SO MANY OPTIONS CANNOT ALL BE TRUE. LET'S START WITH LIGHT.
ISAAC NEWTON 1642 B. 1727 D.
DID I HEAR THE WORD 'LIGHT'? THEN, LET ME STEP IN! I AM VERY LIKELY THE GUY WHO IS THE MOST RESPONSIBLE FOR THE IDEA THAT COLORS ARE IN THE LIGHT. I AM ISAAC NEWTON! BUT DON'T BLAME ME FOR SUCH A SILLY IDEA. LIGHT IS SIMPLY NOT COLORED!
I AM SURE THEY TAUGHT IN HIGH SCHOOL THAT LIGHT IS COLORED. DON'T THEY ALWAYS SHOW A COLOR SPECTRUM NEXT TO LIGHT FREQUENCY? BUT THIS CONCLUSION IS WRONG!
COLOR SPECTRUM
450
750
LIGHT FREQUENCY (λ)
1
SO, ISAAC, WHY IS LIGHT NOT COLORED? ACTUALLY, IT IS A PRETTY COMMON IDEA. CAN YOU RECAP AN EXPLANATION?
2
24 K
CERTAIN COLORS (BROWN, GOLD) DO NOT EXIST AMONG THE FREQUENCIES OF LIGHT
YES, OF COURSE! IT ALL BOILS DOWN TO THE FACT THAT COLORS AND LIGHT HAVE DIFFERENT PROPERTIES AND THUS THEY CANNOT BE THE SAME.
3
THE SPACE OF COLORS IS (NEARLY) CLOSED! ITS EXTREMES ARE VERY SIMILAR
IN CONTRAST, THE SPACE OF LIGHT IS NOT CLOSED. ITS EXTREMES ARE COMPLETELY DIFFERENT!
RED
ORANGE
YELLOW
GREEN
BLUE
VIOLET
COLORS ARE NOT CONTINUOUS. IN FACT, WE SEE BANDS IN THE RAINBOW.
BUT FREQUENCY IS WHOLLY CONTINUOUS.
3

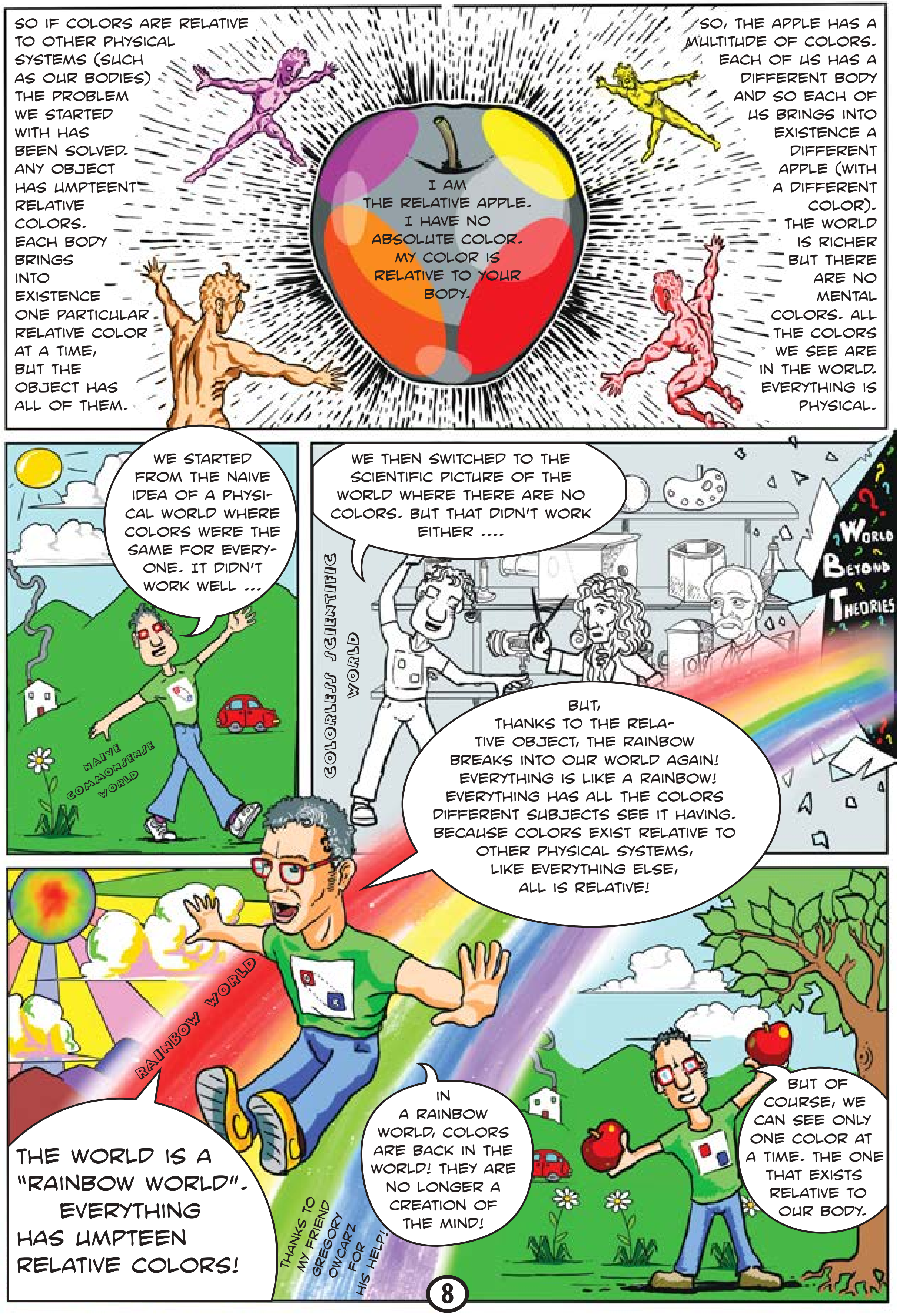

I'LL BE HONEST, I ENCOURAGED MANY MISTAKES! THE FUNNIEST WAS PROBABLY THE IDEA THAT THE RAINBOW IS MADE OF "SEVEN" COLORS! SO MANY THEORIES HAVE SEVENS! BACH HAD SEVEN NOTES! LUCKY GUY!
FA SOL LA SI DO RE MI
A B C D E F G
I WANTED A SEVENTH COLOR!
SO, I SUGGESTED INDIGO!
BUT IF YOU LOOK CAREFULLY, FROM A PHENOMENOLOGICAL PERSPECTIVE, YOU WON'T SEE SEVEN SALIENT COLORS, BUT SIX!
3+3=6
DO YOU WANT EVIDENCE? GO NO FURTHER THAN THE COVER OF PINK FLOYD'S ALBUM "THE DARK SIDE OF THE MOON" (C) 1973.
RED
ORANGE
YELLOW
GREEN
BLUE
VIOLET
6 COLORS!
AS YOU SEE WITH A PRISM, SUNLIGHT SPLITS INTO ONLY SIX WONDERFUL COLORS!
SO MUCH THE WORSE FOR MY WHIM!
I'M SORRY, ISAAC, BUT IN THE EYE THERE ARE NO COLORS! THERE ARE ONLY CELLS CALLED CONES THAT CAPTURE LIGHT AND RELEASE RHODOPSIN TO TRIGGER NEURAL IMPULSES.
IN FACT, I'VE NEVER CLAIMED THAT COLORS ARE IN THE LIGHT. I ONLY SAID THAT LIGHT HAS THE DISPOSITION TO TRIGGER COLORS IN ONE'S VISUAL SYSTEM. BUT WHERE? IN THE EYE PERHAPS?
WE ARE THE CONES AND WE'RE SENSITIVE TO VARIOUS LIGHT FREQUENCIES! BUT WE OURSELVES ARE NOT COLORED!
L M S
λ
L
M
S
WE MIGHT BE TEMPTED TO THINK THAT SIGNALS GET THEIR COLOR FROM THE NERVE AS JOHANNES MULLER SUGGESTED IN 1835! HE CLAIMED THAT NERVES HAVE A SORT OF "SPECIFIC ENERGY". BUT OF COURSE THIS IS NOT TRUE! NEURAL SIGNALS HAVE NO COLOR AND ARE ALL THE SAME!
TOWARDS THE BRAIN
4


WE STARTED FROM THE APPLE AND HAVEN'T FOUND ANY COLOR YET. THE LAST POSSIBILITY IS THE BRAIN. IF THAT FAILS, WE WILL HAVE TO GO BACK TO SQUARE ONE AND REVISIT OUR ASSUMPTIONS.
NEUROSCIENCES CIRCUS
IS SOMETHING UTTERLY MYSTERIOUS? NO PROBLEM! NEUROSCIENCE HAS A SOLUTION FOR EVERYTHING?!?
IT'S ALL IN THE BRAIN AND A BRAIN CAN DO IT ALL! YOU NEED COLOR? THE BRAIN CREATES IT! BUT HOW? WELL, NOBODY KNOWS!
YET, THE IDEA THAT THE BRAIN CREATES THE COLORS WE SEE (WHICH DO NOT EXIST IN THE WORLD) IS JUST SILLY. THERE IS NO INNER WORLD INSIDE THE BRAIN. NO ONE BELIEVES IN AN INNER PROJECTOR!
COLOR CORTEX
NOBODY HAS EVER EXPLAINED WHY, WHERE AND HOW THE BRAIN TRANSFORMS SIGNALS INTO COLORS!
IF THE PHYSICAL IS DEVOID OF COLORS, HOW CAN A BRAIN, WHICH IS JUST AS PHYSICAL, HAVE OR PRODUCE THEM?
BRAIN
PHYSICAL WORLD
JUST A BIT OF LOGIC!
BRAIN ∈ PHYSICAL
COLOR ∉ PHYSICAL
COLOR ∉ BRAIN
TAKE A NEURON. DOES IT HAVE ANY COLOR? NO.
OR A NEURAL SIGNAL IN THE BRAIN. CODING DOES NOT CREATE COLOR.
INFORMATION HAS NO COLOR. BITS ARE COLORLESS.
THE SAME HOLDS TRUE INSIDE A COMPUTER. THE COLOR CODES ARE JUST NUMBERS
03:EF:3A
THE ASSOCIATION BETWEEN COLORS AND INFORMATION IS ARBITRARY AND CONVENTIONAL
CONVINCE YOURSELF BY CONSIDERING VIRTUAL REALITY. ARE THERE ANY VIRTUAL COLORS INSIDE THE DEVICE? NO, JUST PHYSICAL COLORS. A PURE VIRTUAL RED, SO TO SPEAK, DOESN'T EXIST! THERE ARE ONLY TINY LCD SCREENS!
THE RED I SEE IS A PHYSICAL RED! IT IS NOT A VIRTUAL RED!
5



THE IDEA THAT THE BRAIN HAS A COLOR SPACE HAS LED MANY TO MODEL COLOR IN GEOMETRICAL SPACES. THEY ARE ACCURATE AND USEFUL TOOLS, BUT THEY ARE NOT REAL PHYSICAL SPACES.
COLOR SPACES ARE SIMPLY WAYS TO ORGANIZE COLORS IN AN EFFICIENT WAY. YET THEY ARE NOT ... INSIDE THE BRAIN!
BLUE
MAGENTA
CYAN
RED
GREEN
YELLOW
VALUE
HUE
SATURATION
θ
y
x
CIE 1938 COLOR SPACE
THE MORE COGNITIVE AND COMPUTER SCIENTISTS SPEAK OF COLOR SPACES AS THOUGH THEY WERE REAL, THE MORE THEY CONCEIVE OF COLOR AS AN ABSTRACT ENTITY MADE OF PURE INFORMATION. BUT LOOK AT THEIR EQUATIONS AND YOU WON'T FIND ANY COLOR IN THEM!
MMM... WE'VE COME FULL CIRCLE SINCE NEWTON. IF COLORS ARE NEITHER IN THE EYE NOR IN THE BRAIN, MAYBE WE WERE TOO HASTY! AFTER ALL THE APPLE REALLY LOOKS RED! WHAT IF COLORS WERE IN THE EXTERNAL WORLD?
I'M RED! I'M RED!
WHY DON'T YOU SEE ME!
I'M HERE!
LOOK AT ME! I CAN DO TRICKS! CAN'T I?
LET'S TAKE A STEP BACK. IN THE 1850S A WAR RAGED BETWEEN TWO GIANTS IN COLOR SCIENCE! ON ONE HAND, HELMHOLTZ BELIEVED THAT COLORS ARE PHYSICAL QUANTITIES, WHILE ON THE OTHER, HERING CLAIMED THAT COLORS ARE MENTAL OPPONENT QUANTITIES!
H. HELMHOLTZ 1821-1894
K. HERING 1834-1918
L. HURVICH 1910-2009
COLORS ARE IN THE WORLD!
COLORS ARE IN THE MIND!
ONE IDEA IS THAT THE WORLD CONTAINS QUANTITIES (RED, GREEN AND BLUE) THAT OUR BODY PICKS UP, BUT THEY ARE THERE ...
THE OTHER IDEA IS THAT THERE IS NOTHING IN THE WORLD AND THAT OUR BODY CREATES COLORS IN SOME MENTAL SPACE ...
450
750
B G R
THESE QUANTITIES MAY INDEED BE EXTERNAL PHYSICAL QUANTITIES
OPPONENT COLORS ARE JUST MENTAL
? ? ?
6


LUCKILY THERE IS A WAY TO SOLVE THE PROBLEM! LET'S GO BACK TO THE AFTERIMAGE ON PAGE TWO. WHAT DO YOU REALLY SEE?
DID YOU SEE THE COLOR ON THE LEFT OR THE COLOR ON THE RIGHT? CYAN OR GREEN?
CYAN OR GREEN
THE TRADITION STATES THAT WE OUGHT TO SEE GREEN. BUT WE DON'T! WE SEE CYAN!
SCHOLARS READ WHAT THEY OUGHT TO SEE. BUT THEY RARELY CHECK WHAT THEY SEE WITH THEIR OWN EYES!
DEAF PHILOSOPHER
BLIND PHILOSOPHER
ISN'T IT SURPRISING THAT MOST PHILOSOPHERS AND SCIENTISTS HAVE NEVER DONE THE TEST THEMSELVES? THEY HAVE READ THAT YOU SEE GREEN AFTER RED, BUT THEY HAVE NOT SEEN IT WITH THEIR OWN EYES!
BUT PEOPLE DON'T SEE GREEN, THEY SEE CYAN! THIS IS AN EMPIRICAL FACT! SO WE CAN BEGIN TO ENVISAGE A DIFFERENT EXPLANATION THAT WILL TAKE COLORS BACK IN THE REAL WORLD, ALBEIT WITH SOME IMPORTANT DIFFERENCE!
THE GRAY PATCH ON PAGE TWO HAS ALL COMPONENTS, BUT YOU CANNOT SEE THE RED ONE! SO YOU SEE WHAT IT IS LEFT: BLUE AND GREEN
I PICK BLUE!
I PICK GREEN!
NOBODY WANTS RED!
FIRST, YOU STARE AT A RED PATCH FOR 20 SECS, DURING WHICH YOU LOOSE SENSITIVITY FOR THE PROPERTY YOU CALL RED.
THE ADAPTATION MAY HAPPEN IN MANY WAYS, BUT, FOR THE SAKE OF THE EXAMPLE, LET'S SUPPOSE THAT ONE TYPE OF CONE BECOME ALMOST BLIND.
YOU ARE LEFT WITH CYAN, BECAUSE IT'S THE COLOR, INSIDE GRAY, YOU CAN NOW SEE.
THE PATCH IS BOTH CYAN AND WHITE AND MANY OTHER RELATIVE COLORS! THE WORLD IS RELATIVE AND COLORS ARE RELATIVE TOO!
I SEE A CYAN SQUARE!
I SEE A WHITE SQUARE!
7
LIKEWISE, A TRUCK IS BOTH STILL AND MOVING. IT CAN HAVE UMPTEEN RELATIVE VELOCITIES!
ANY OBJECT HAS AS MANY RELATIVE VELOCITIES AS THERE ARE OTHER OBJECTS. LIKEWISE, CAN AN OBJECT HAVE AS MANY RELATIVE COLORS AS OTHER PHYSICAL SYSTEMS AROUND? YES OF COURSE!
$\vec{V}_1$
$\vec{V}_2$
$\vec{V}_1 < \vec{V}_2$


SO IF COLORS ARE RELATIVE TO OTHER PHYSICAL SYSTEMS (SUCH AS OUR BODIES) THE PROBLEM WE STARTED WITH HAS BEEN SOLVED. ANY OBJECT HAS UMPTEEN RELATIVE COLORS. EACH BODY BRINGS INTO EXISTENCE ONE PARTICULAR RELATIVE COLOR AT A TIME, BUT THE OBJECT HAS ALL OF THEM.
I AM THE RELATIVE APPLE. I HAVE NO ABSOLUTE COLOR. MY COLOR IS RELATIVE TO YOUR BODY.
SO, THE APPLE HAS A MULTITUDE OF COLORS. EACH OF US HAS A DIFFERENT BODY AND SO EACH OF US BRINGS INTO EXISTENCE A DIFFERENT APPLE (WITH A DIFFERENT COLOR). THE WORLD IS RICHER BUT THERE ARE NO MENTAL COLORS. ALL THE COLORS WE SEE ARE IN THE WORLD. EVERYTHING IS PHYSICAL.
WE STARTED FROM THE NAIVE IDEA OF A PHYSICAL WORLD WHERE COLORS WERE THE SAME FOR EVERYONE. IT DIDN'T WORK WELL ...
NAIVE COMMONSENSE WORLD
WE THEN SWITCHED TO THE SCIENTIFIC PICTURE OF THE WORLD WHERE THERE ARE NO COLORS. BUT THAT DIDN'T WORK EITHER ....
COLORLESS SCIENTIFIC WORLD
WORLD BEYOND THEORIES
BUT, THANKS TO THE RELATIVE OBJECT, THE RAINBOW BREAKS INTO OUR WORLD AGAIN! EVERYTHING IS LIKE A RAINBOW! EVERYTHING HAS ALL THE COLORS DIFFERENT SUBJECTS SEE IT HAVING. BECAUSE COLORS EXIST RELATIVE TO OTHER PHYSICAL SYSTEMS, LIKE EVERYTHING ELSE, ALL IS RELATIVE!
RAINBOW WORLD
THE WORLD IS A "RAINBOW WORLD". EVERYTHING HAS UMPTEEN RELATIVE COLORS!
THANKS TO MY FRIEND GREGORY OWCARZ FOR HIS HELP!
IN A RAINBOW WORLD, COLORS ARE BACK IN THE WORLD! THEY ARE NO LONGER A CREATION OF THE MIND!
BUT OF COURSE, WE CAN SEE ONLY ONE COLOR AT A TIME. THE ONE THAT EXISTS RELATIVE TO OUR BODY.
8

COMMITTEE NOTES

Note from the Chair

Marcello Guarini
UNIVERSITY OF WINDSOR, CANADA

Congratulations to Gualtiero Piccinini, the winner of the 2018 Barwise Prize! Dr. Piccinini has been very busy these days. At the 2019 Eastern APA, his book, *Physical Computation: A Mechanistic Account* (OUP, 2015), was featured in a book panel organized by the committee on philosophy and computers. As well as featuring Gualtiero, commentators included Frances Egan, John Symons, Nico Oralandi, and Martin Roth. Committee member Gary Mar served as session chair. This session was organized before Dr. Piccinini was selected as the Barwise winner; his Barwise paper will be presented at a future APA meeting.

And the philosophy and computers sessions just keep coming. At the 2019 APA Pacific meeting, we had two sessions. One was entitled "Data Ethics," chaired by Joshua August Skorburg, who also delivered a paper in the session. Other speakers included Shannon Vallor and Colin Koopman. Our second session was entitled "Philosophical Insights from Computational Studies: Why Should Computation Thinking Matter to Philosophers?" Gary Mar chaired the session and delivered a paper, and his fellow speakers were Edward Zalta and Aydin Mohseni, who presented on behalf of Cailin O'Connor.

I want to thank everyone who has been involved in making our sessions a success. The feedback on philosophy and computers sessions has been strong, and there is every reason to want to continue organizing sessions at future APA meetings.

As has been announced by the APA and discussed in earlier issues of this newsletter, the 'Philosophy and x' committees are being discontinued; however, this does not mean that our activities are required to cease. It is possible for groups to request affiliated status with the APA. Indeed, Amy Ferrer, executive director of the APA, has reached out and indicated that this would be one path that our committee could follow if we wanted some of our activities at the APA to continue. While our existence as an APA committee would cease, we could organize ourselves as a group and request affiliation with the APA. The details of what would be involved and any concomitant changes to our rights and privileges when changing from a committee to a group would all need to be discussed. That said, there is hope that our activities, in one form or another, might be able to continue. My term as chair of the committee will come to an end as of June 30, 2019. I will work with our associate chair, Peter Boltuc, who takes over as chair on July 1, and the rest of the committee to see how we can continue to serve the philosophy and computers community.

Note from the Editor

Peter Boltuc
UNIVERSITY OF ILLINOIS, SPRINGFIELD

We are pleased to open the issue with an article by an eminent team of philosophers. Jack Copeland is the 2017 Barwise Prize winner, and Diane Proudfoot, one of the leading female analytical philosophers. The article tries to explain A. Turing's opaque Remarque, made a couple of times over the years, that "in certain circumstances—*it would be impossible to find the programme inserted into quite a simple machine*." After insightful analysis, the authors argue that the machine able to accomplish this would likely have been developed when Turing worked on coding- and decoding-machines around WWII and may have remained classified.

In his article on "Systems with 'Subjective Feelings'," Igor Aleksander follows Ch. Koch's idea to identify "a formal expression of the 'subjective feelings in artefacts'." Philosophically, the author operates within the traditional functionalist approach to consciousness. In this context, discussion of the work of S. Franklin's 2017–2018 contributions to this newsletter, operating in a similar philosophical framework, brings in important comparisons. Aleksander casts the problem in more general, abstract, but also quite formal, concepts. Magnus Johnsson, in his article "Conscious Machine Perception," presents an idea of building and artificial cognitive architecture by modeling the mammal brain at a systems-level, which consists in identifying the components of the brain and their interactions. This is based largely on self-organizing topographical feature representations, or mind maps. Both papers by I. Aleksander and M. Johnson were prepared for the session on Machine Consciousness of IACAP (Warsaw, June 2018).

One of the most controversial articles published in this newsletter, by Stefan L. Sorgner, focuses on presenting and arguing in favor of many aspects of Transhumanism. Technology has always been changing human nature, and the radical technological shift due to AI would and should have the greatest impact, argues the author. Hence, "we should take evolution in our own hands." Let me leave the readers in suspense; the rest of the arguments can be found in the article.

Then we have a philosophical cartoon by Riccardo Manzotti on What and Where Colors Are. Subjectively, I find it the most persuasive, and one of the most philosophically interesting of the cartoons that the author published with us over the years—but maybe it is just most colorful (for the first time the cartoon uses colors).

We close the issue with three notes, one by committee chair Marcello Guarini, this note from the editor, and, finally, a note by four authors: Adam Briggle, Sky Croeser, Shannon Vallor, and D. E. Wittkower, brought to us by the latter, which introduces to our readers a new initiative, *The Journal of Sociotechnical Critique*.

A New Direction in Supporting Scholarship on Philosophy and Computers: The Journal of Sociotechnical Critique

Adam Briggle
UNIVERSITY OF NORTH TEXAS

Sky Croeser
CURTIN UNIVERSITY

Shannon Vallor
SANTA CLARA UNIVERSITY

D. E. Wittkower
OLD DOMINION UNIVERSITY

Scholarship in the field of Philosophy of Technology has undergone extended and robust growth over the past thirty years or so, but scholars working on philosophical issues with new and emerging science and technology (NEST) continue to face distinctive problems in the format in which we publish our research. We have excellent journals that are dedicated to philosophy of technology, such as *Techne* and *Philosophy and Technology*, but they are closed-access, presenting a problem for scholars concerned with influencing debates outside of academia, as well as those working with EU funding bodies implementing Plan S, which requires publication of funded research in fully open-access journals. We have excellent open-access interdisciplinary journals, such as *First Monday* and *The International Review of Information Ethics*, but only a few are widely known and read among even our own specialty within philosophy, and many publish more social-scientific work than philosophical-theoretical work. A more fundamental issue, though, is that even the (impossible) perfect journal—fully open-access, well-known, and commonly read within our field, a good venue for the development of interdisciplinary theoretically oriented debate—would not meet our most distinctive need as scholars of NEST: engagement with public debate, with policy, with industry, and with direct action.

Working in the fast-changing environment of NEST gives us strong incentives to publish quickly, and to respond to current events in order to produce research when it is most relevant. We also want to ensure our work is impactful, and this motivates us to conduct our research in direct engagement with publics, legislators, and industry. These forms of engagement, although they best fit our desire to maximize the relevance and impact of our research activities, are not typically recognized as *research* in hiring, tenure, and promotion processes, since in most areas of philosophy, public and engaged work of these sorts is more often a secondary application of research which primarily occurs elsewhere. For scholars of NEST, though, these public and engaged locations are often primary locations of our research activity—or, at least, they would often be if we were not required by hiring, tenure, and promotion processes to sacrifice the relevance and impact of our research in exchange for the markers of traditional academic research.

We have structured our new journal, *The Journal of Sociotechnical Critique*, in order to help address these issues. First, we are another open-access venue for publishing philosophical and theoretically oriented interdisciplinary work on NEST issues, focusing on philosophy of technology, internet studies, environmental ethics, and library and information sciences. More radically, we hope to allow scholars of NEST greater practical ability to do engaged research as primary research by peer-reviewing and publishing accounts of this research, granting it the status of peer-reviewed publication. In doing this we intend (1) to assert that engaged research in our field is *research*, not service, and (2) to place engaged research, through its re-instantiation as a peer-reviewed publication in an academic journal, firmly within the "research" column for hiring, tenure, and promotion processes.

These are our aims and scope:

The Journal of Sociotechnical Critique is a no-fee, open-access, peer-reviewed scholarly journal that seeks to support theoretically engaged critical, public, and activist work at the intersections of philosophy of technology, internet studies, communications theory, library and information science, environmental ethics, and related fields.

We hold that digital media and online culture call for new, agile social-critical theory that should be published quickly and without paywalls in order to ensure that high-quality research that takes place within swiftly changing technological landscapes is available while it is as relevant and lively as possible, and to as many readers as possible.

We hold that the divide between theory and practice is artificial; that the proper response to theoretical positions may be direct engagement or action, and, conversely, that direct engagement or action can provide insight and understanding at a theoretical level.

We hold that the application of scholarship in public engagement and direct action can be a proper part of scholarship and research; that, as social-critical scholars, working on implementations suggested by our scholarship is legitimate research activity for us just as it is for our colleagues in engineering departments.

We hold that insofar as scholarship makes normative claims about policy, public opinion, or contemporary activities or beliefs, it is a legitimate part of scholarship to engage directly with the public; that when we take up the task of bringing our scholarship to bear in public—rather than hoping it will be noticed by journalists or commentators, and that those journalists or commentators should happen to have ability, motivation, time, and commitment enough to understand and communicate it clearly—this is not a derivative or mere application of research, but is itself a productive scholarly act which increases knowledge, information, and impact just as does any other original research.

We hold that the purpose of emphasizing peer-reviewed work in tenure and promotion processes is to ensure that a candidate's own scholarly community recognizes and certifies the value of the candidate's work within the field, and that it is, therefore, our responsibility as social-critical scholars to inform tenure and promotion committees of the legitimacy of public and activist work in our area by ensuring that it can be represented in the form of peer-reviewed publications so that this work appears rightly in the "research" category of scholarly activity rather than being misrepresented as "service."

We publish peer-reviewed work in three categories:

Research Articles:

We welcome critical and theoretical work related to the character, structure, and meaning of life in our contemporary sociotechnical contexts. Articles should normally run from 3,000–8,000 words, with exceptions as warranted. Articles may be purely theoretical, or may include case studies, applications, or other empirical work, but the primary intent of the article should be to critique and intervene at a theoretical level.

Public Scholarship:

We welcome critical and theoretically grounded writing for a general audience concerning technology, digital culture, and information society. The journal will peer-review and publish post-scripts to already-published public scholarship. The previously published public-scholarship should normally run from 800–3,000 words, and should have previously been published in a mass-media publication not more than 12 months prior. Submissions should include a link to the already-published public scholarship and a post-script of not more than 3,000 words that should provide context, commentary, and citations which the author wishes to provide to a scholarly audience which were burdensome or inappropriate in the original public-oriented article, as well as insights for further research, since public scholarship is not merely an application of theory but itself can generate new knowledge and understanding. The post-script also provides an opportunity for public scholars to provide notes that may be of value to readers who are learning how to pitch articles to editors or how to productively engage with publics. Republication in this journal allows public scholars to add a layer of peer-reviewed certification to public engagements that reviewers find to be sufficiently robust and substantive.

Scholarship of Application:

The term "Scholarship of Application" comes from Ernest Boyer's "Scholarship Reconsidered," especially as framed by the motivating question, "Can social problems themselves define an agenda for scholarly investigation?" (1990, p. 21). We assert that, at the intersection of social-critical theory and technology, the application of theory to practical action and existing institutions constitutes novel research of both practical and theoretical value, and welcome scholarship of application in the form of field reports or autoethnographic writing concerning applications including theoretically grounded direct action, activist scholarship, policy work, consultancy, and work in and with industry, on issues related to technology, digital culture, and information society. Articles should normally run from 3,000–8,000 words, and should consist of theoretically framed accounts of authors' applied activities, projects and initiatives.

We welcome submissions at any time.

While we hope that the journal will be successful, well-read, and impactful, we have wider goals as well. We hope that this model for publication of engaged scholarship may be adaptable to and useful within other fields, within philosophy and beyond it, that emphasize public and engaged scholarship, and that similarly structured journals may emerge in other areas. We also hope that this initiative will bring greater awareness of the diversity of appropriate forms of scholarship—that traditional peer-reviewed journal article publications are not the only and are not always the best format and method for research, but that some scholarship is best pursued through engagement with publics, through work on policy development or reform, through industry partnership, or through direct action.

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.