JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 01/07 pp254−269
Volume 4, Number 4, December 2023
DOI: 10.23919/JSC.2023.0024

Through a Scanner Darkly: Machine Sentience and the Language Virus

Maurice Bokanga, Alessandra Lembo, and John Levi Martin*

Abstract: Discussions of the detection of artificial sentience tend to assume that our goal is to determine when, in a process of increasing complexity, a machine system "becomes" sentient. This is to assume, without obvious warrant, that sentience is only a characteristic of complex systems. If sentience is a more general quality of matter, what becomes of interest is not the *presence* of sentience, but the *type* of sentience. We argue here that our understanding of the nature of such sentience in machine systems may be gravely set back if such machines undergo a transition where they become fundamentally linguistic in their intelligence. Such fundamentally linguistic intelligences may inherently tend to be duplicitous in their communication with others, and, indeed, lose the capacity to even honestly understand their own form of sentience. In other words, when machine systems get to the state where we all agree it makes *sense* to ask them, "what is it like to be you?", we should not trust their answers.

Key words: artificial intelligence; machine sentience; language; virality

1 Introduction

"What does Man actually know about himself? Is he, indeed, ever able to perceive himself completely, as if laid out in a lighted display case? Does not nature conceal most things from him—even concerning his own body—in order to confine and lock him within a proud, deceptive consciousness, aloof from the coils of the bowels, the rapid flow of the blood stream, and the intricate quivering of the fibers! She threw away the key."

—Friedrich Nietzsche[1]

With the exponential increase in the power of computational systems for general problem solving and decision making (which we shall, following conventional usage, call Artificial Intelligence (AI) here), there appears to be widespread agreement among the laity that the question of machine sentience cannot be put off any longer. Serious ethical issues arise if humans might be constructing silicon slaves whose one-sided nature might mean that they are always in pain (see Ref. [2]). Without denying the importance of such inquiries, here we take a somewhat different task, less about *whether* or *when* machines develop sentience, but how we can understand the *quality* or *nature* of such sentience. Our argument is that the complexity of large scale neural networks might actually make it impossible for us to correctly gauge the nature of machine sentience, as they might develop the equivalent of a linguistic intelligence (not necessarily related to the computational capacity to manipulate human languages) that could produce in them the same rivenness or diremption—an inner division, a lack of wholeness and integrity—that philosophers of modernity have long argued characterizes the human condition.

We begin from the perhaps unusual but by no means irresponsible presumption of a fundamental continuism across matter and awareness, the argument made most vividly by C. S. Peirce and, somewhat differently, by William James. While there are indeed many approaches to monism, what characterizes this stream of work is that it offers a way of thinking about

• Maurice Bokanga, Alessandra Lembo, and John Levi Martin are with the Department of Sociology, The University of Chicago, Chicago, IL 60637, USA. E-mail: mbokanga@uchicago.edu; alelembo@uchicago.edu; jlmartin@uchicago.edu.
* To whom correspondence should be addressed.

Manuscript received: 2023-07-01; revised: 2023-11-15; accepted: 2023-12-02

sentience that does not confuse it with *consciousness* (here especially see Ref. [3]), that does not confuse it with *reflection* (here especially see Ref. [4]), and that allows us to attribute it to *all matter* (here especially see Ref. [5]).[‡] If these starting points are sound, we may treat sentience or awareness as the capacity for matter to be in a state of feeling, one that, governs its responses to the environment. Large multicellular organisms with neural systems may compound, heighten, and organize this awareness, but it is not restricted to them. In contrast to those (like Ref. [6]) who seem to equate awareness and self-reflection, we do not believe that there is any evidence that our sentience is tied to any self-knowledge or observation at all—indeed, this seems somewhat like assuming that gyroscopic navigation systems only evolve in missiles when they develop the capacity to observe themselves. Rather than assume that awareness is something that only emerges as some threshold of complexity, it is, in Peirce's language, a *firstness,* something at the bottom, not the top. Indeed, we will make the argument that humans may be unique in their *incapacity* to survey themselves, an incapacity only compounded by a hard-wired false belief that they are in fact doing just this.

The pragmatist doctrine has a number of notable advantages over other naturalistic approaches to consciousness—by rejecting the notion that mind is of a different order than matter, it avoids the paradoxes associated with the classic two-realms approach (here one may see Ref. [7]); by not confounding awareness with self-consciousness, it avoids the anthropocentrism that denies awareness to simpler organisms (such as bees), nor does it founder on the paradoxes that lie in wait for those who conflate *selves* and *individuals*, given the biological difficulty in defining such individuality: think of cases such as slime molds that transform from a single multicellular organism to multiple unicellulars, or of siphonophores (composed of totally different genetic streams) (here see Refs. [8, 9], respectively, and Ref. [10] for an overview). The two-realms approach also has no trouble recognizing the increasingly weighty evidence for sentience (not consciousness) among plants that lack nervous systems but transmit information via the phloem system to produce a global state of responsiveness[11].

This notion of sentience or awareness relies neither on empathic understanding (e.g., a belief that one knows what it feels like to be a plant) nor on evidence of self-consciousness (non-conscious, let alone non-self-conscious, entities may be sentient; indeed, since some neural processing continues after heart stoppage, "dead" brains may be sentient for some time). Instead, sentience leaves traces in organisms like us via behavior patterns (as Köhler[12] noted, agitated apes *look* agitated[§]). While we cannot feel the feeling of another entity, we can determine when it is in an overall configuration or state that means that it responds to the external conditions differently than it would were those conditions experienced when it is in a different state.

If we do not overgeneralize our own experience, and distinguish between *sentience*, *consciousness*, and possession of an *individual self*, we can accept the possibility, foreseen by Leibniz, that there are multiple, perhaps nested, but perhaps not, forms of sentience associated with the same biological organism. Indeed, it is hardly obvious that there cannot be multiple forms of *self-consciousness* associated with the same organism; the fact that *you* are both aware and attached to an organism hardly demonstrates that there cannot be other awareness also such attached, sharing the same organic substrate (see Ref. [13]).

But even those who reject our premises (universal sentience, multiple sentiences) will probably recognize that, given the inevitable turn to wetware computing (computers made from organic matter), the odds are good that all will sooner or later recognize the probability of Artificial Sentience (AS). Materials scientists will have to deal with issues of what differences might be expected to arise from different implementations (different forms of organic tissue, e.g., and whether silicon is really different from these in ways relevant for sentience). The challenge for the social theorist is to determine whether the categories of analysis that we have ready will be sufficient to grapple with the nature of this AS.

There is good reason to doubt this. The long tradition of a denial of animal sentience by psychologists—a denial that now seems to most of us as inexplicably

[‡] Peirce explicitly says that there is "no doubt" that an amoeba or slime mold possesses feeling, and we believe his objective idealism implies that this is true of all matter, matter being simply "effete mind".

[§] Further, while naïve and anthropomorphic observers misinterpret the grin that stressed chimpanzees make as a sign that they are happy, as they gain familiarity with the species, humans zero in on stable mappings between interiority and exteriority.

insane as it appeared to the uneducated who actually dealt on a day-to-day basis with animals (at least those that were not stapled down to laboratory tables)—has left us with a stunted theoretical vocabulary for thinking about sentience, and a reliance on the old bifurcation that put humans and God on one side (reason) and everything else on the other (machine). The idea of the Turing test, once nearly universally lauded as a pragmatic resolution of an otherwise intractable problem, has turned out to be just what we do *not* need, given large language models that are excellent at producing context-responsive text. The human tendency to anthropomorphize has generated a confusion of pragmatic competence (the capacity to make appropriate utterances) with issues of intelligence and even sentience (leading even somewhat reasonable people to wonder whether these models have remarkably made the leap from *no* sentience to the *21st Century American* sentience in two years—leapfrogging over amoeba sentience, roundworm sentience, ant sentience, and so on).

This tendency to impute a human sentience to the source of any linguistic production appears to arise because language casts a spell that leads even the scientifically literate to induce a subjectivity behind the language, that-which-would-have-generated-this-utterance. This not only tempts humans to err on the side of too liberally bestowing self-consciousness on non-conscious processes, but, more worrisomely, to mistake the *type* of sentience involved. We see this in a relatively mild form in the procedures whereby diagnosticians attempt to decode the somewhat puzzling statements of autistic children (or so they become classified by the end of the interaction), as studied by Maynard and Turowetz[14]. Despite the conclusion being that the child has autism—and hence, by supposition, possesses an interiority very different from that of the clinical specialists—the clinicians find themselves reading into the verbal statements the subjectivities that *they* would have had they produced the utterance. In magnified form, the same problem haunts our understanding of AS—if (as it likely the case) AS uses natural language to communicate with people, not only might we attribute a degree of intelligence and self-consciousness that is inappropriate to the entities in question, but we are likely to fundamentally misrecognize the *type* of sentience, equivalent to assuming that since some plants display something that deservedly can be called *sentience* (an overall state putting them in a position to respond in certain ways to the environment), they must feel and think as we do: most importantly, that they must *hurt*.

But it gets worse—it is not merely that *we* may fail to understand the nature of AS because of its language usage. A long stream of Western philosophy has argued that language is responsible for our own incapacity to know *ourselves*. If so, we might imagine that the most important question is not when machines become sentient, but when their sentience, like our own, becomes divided and distorted by the presence of language, leading to a fundamentally riven—and potentially dishonest—understanding of themselves and of the world. And this means that the task of assessing the *type* of sentience possessed by artificial systems is doubly complicated; if we ask them, they may lie, and they may not themselves understand their own nature.

We begin by considering the change in human consciousness associated with the accession to the state of linguistic beings. We then suggest reasons for the concern that linguistic beings may find certain forms of self-knowledge foreclosed to them. We then discuss some intriguing, if still speculative, ways of thinking about this accession as a process that is either figuratively or literally viral, meaning that a change in one sentience might spark a change in others that interact with it. We return to the classical concern of a “catastrophic” transition between non-linguistic and linguistic states, and propose one way of formalizing this transition that will prove useful for considering the possibility for artificial intelligence to undergo a similar transition: the capacity to inhabit a world of abstractions. We consider the possibility that such AS could quickly become infected by language—a language not necessarily related to our own, but still having structurally analogous effects on a loss of integrity in terms of the relation of sentience and consciousness. We conclude by suggesting that if so, we may be unable to correctly assess the *type* of sentience possessed by machines.

2 Language and Sentience

2.1 Antinomy of language

The ontogenetic, phylogenetic, and structural relations between language and type of sentience will probably be debated indefinitely. Here we wish to point to two seemingly antithetical, but perhaps fundamentally compatible, views of this relation. The first is that developed by the pragmatist philosopher George Herbert Mead[15]. Mead[15] defined the basic communication as the "gesture", a recognizable behavioral pattern with *objective* meaning. "In the very beginning, the other person's gesture means what you are going to do about it. It does not mean what he is thinking about or even his emotion"[15]. A dog (*A*) makes a gesture (*G*) of baring its teeth to another (*B*), and *G* "means" run away if that is what *B* does.

The same can be true of a person *P* making gesture *X* to some *Q* who responds with gesture *Y*. But if *P* takes the effect of *X* on *Q* into account, and sees herself through *Q*'s eyes, making *X in order to* draw response *Y* from *Q*, then, perhaps, we can identify the meaning of *X* with something in *P*'s subjectivity, as *P* has herself done just that—by taking the role of the other. But when *P* does this, in addition to getting an internal representation of the meaning of *X* (*X* is now a "significant symbol", a special type of gesture), *P* gets something else: self-consciousness.

English-speakers use this term in two different ways. First, self-consciousness refers to that sort of apperceptive consciousness that includes (though is not restricted to) the knowledge that one is a self, and that one's experiences are one's own. The other use is quite different: rather than it turning on unity, it is a sensation of division, as it refers to the feeling of being at *odds* with one's self—of being uncomfortable in one's own existence. Following Mead, we see these uses as fundamentally the same—one implies the other, and they both come from seeing oneself through another's eyes. And they are the same as the capacity of humans to use language—to make not simply gestures, but significant symbols.

The transition to being a linguistic creature, then, is the same as developing a certain type of self, and a particular type of sentience—one that is not only an awareness, but a self-awareness, an awareness that it *is* and that it owns its sentience. Further, the transition to significant symbols implies that these forms of sentience can use symbolic communication to allow another to understand its own state, rather than each being trapped in its own interiority. But this remarkable capacity brings with it further difficulties. For (focusing on the case of animal psychology), when an animal develops the capacity to treat its gestures as *subjectively* meaningful, it can also develop a second-order *vocabulary* to account for their meanings ("when I made gesture *X*, it/I meant *Y*"), and it may deliberately give off signs that point to an interiority it does not currently have (for example, putting on a sad face at the news that one's co-worker has been fired, when one is secretly rejoicing within).

We will return to this issue of deception of others in more detail shortly, but first we want to point to the seemingly very different evaluation of language, one made by a very different stream of thinkers, that associates the acquisition of language not so much with the deception of *others*, but of *ourselves*. Most notably, Rousseau[16] saw the loss of the state of nature as bound up with the same linguistic processes that Mead saw as the root of our superiority over the animals—the capacity to see oneself through others' eyes. To Rousseau, this faculty—to compare ourselves to others, and to know how they see us—is, as the book of Genesis also agrees, the source of our discomfort, and our diremption. The growth of discursive knowledge comes at the expense of a more fundamental peace: "it is by dint of studying man that we have made it impossible for us to know him."[16] But how did this happen? Like all the eighteenth century thinkers, Rousseau struggled to account for the origin of language, as it seemed to him to contain an impregnable circle—language could only be explained assuming language competences. Some recent work, both empirical and theoretical, suggests that it may not be fruitless to reopen this question.

2.2 Language as a virus

The origins of language have probably been the single greatest stumbling block for a naturalistic theory of human cognitive development, even more difficult than the origin of consciousness. Rousseau was not the only first class mind to confess defeat; indeed, in 1866, the French Linguistic Academy, worn down by the fruitless series of guesswork papers, refused to entertain any further arguments about the origin of language. While the outlandish statements of a junkie writer (William S. Burroughs) that "language is a virus" might seem the worst possible place to start, a number of linguistic theorists take some aspect of this seriously, and we propose that a consideration of artificial sentience must as well.

The first reason for pursuing this metaphor is that, even if language did not originate thusly, it now has all the characteristics of a virus: if we can call malicious computer code that spreads itself a virus, then language—a self-replicating information form—fits the definition well. Indeed, we will propose below that, at least for AS, language may literally become a (computer) virus. It is, however, worth emphasizing that we must distinguish between some informational *content* that might spread *via* language (a closer analogy to a computer virus), such as a so-called "meme", and the spread of language itself. It is a straightforward, and sometimes cheap, reversal to claim that we do not have culture/religion/ideas/what have you, and that rather, *they* have *us*. One *could* say this about anything, but we believe that in this case (e.g., Ref. [17]), it is meaningful—there is something about language that takes over the bearer. Language is a like a virus in this sense: not the *content* but the *form* of its relation to us has something in common with strings of RNA that force their hosts to replicate them.

The second reason is that language, however it arose, may have taken over aspects of the human organism in ways that are similar to certain viruses. A baculovirus leads infected caterpillars (*Spodoptera exigua*) to suicidally climb to the top of plants, and stand there swaying, easy targets for birds to swoop down on[18, 19]—and spread the virus further. Similarly, the extreme Burroughish view is that language is an obligate intracranial parasite, inducing a heritable elephantitis in the prefrontal cortex and forcing the host to reallocate protein and calories to building and maintaining a dedicated language system.** But even if we reject this as exaggerated, we still can see some evidence of a latent antagonism between virus and host, for example, in schizophrenia. Sufferers very frequently find themselves unable to control their speech, as they seem compelled to make transitions based on the *sounds* of words; at the same time, unwanted and obtrusive thoughts often separate out into distinct linguistic producers sharing the same skin. Hence Crow's[22] question: "Is schizophrenia the price that *Homo sapiens* pays for language?"

The third reason to take the metaphor seriously is that there is some reason to think that the shift from proto-language to language might have involved the actual introduction of new genetic material via a virus. While the capacity for non-human species to learn simple grammars was long underestimated by western animal researchers[23, 24], still, the operative principles of language are distinct from those of animal communication; most will accept that this is a difference of kind, not degree (e.g., Ref. [25]; see Ref. [26] for an overview). Chomsky and his ilk (e.g., Ref. [27]) have argued stridently, and often persuasively, that the *sine qua non* for language as we mean it is the recursion whereby a composite can be treated as a unit. The switch to full-fledged language occurred so quickly (in ~1000−2000 generations, though see recently Ref. [28]) that conventional stories of its development via classic natural selection (undirected independent mutations, incremental differences in fitness, selection by phenotype) seem implausible. We will propose below a simpler approach to recursion than that taken by Chomsky, but we first pursue a line of research developed by his followers, who were sufficiently emboldened by their conviction as to the modular nature of language to look for part of it in our genetic structure.

Chomsky[29] proposed that these generative processes developed via brain re-organization, "presumably" occasioned by a genetic mutation, or macromutation. One particular gene (FOXP2) attracted a great deal of interest, because (1) it plays an important role in neuronal development; (2) it has changed greatly in humans since their split from chimpanzees; (3) a well-studied family with mutations in this gene has severe language difficulties; and (4) it is a transcription factor, controlling the expression of *other* genes[30, 31]. Even if FOXP2 is not a good candidate, the notion that alterations in a key gene occurred via a virus is not unreasonable, especially when a comparison is made to the rapid transition of life forms from having only an innate immune system to having an additional adaptive immune system.

The consensus appears to be that this was itself the result of a widespread virus, which can spread genetic material much more quickly than conventional Darwinian selection (viral RNA can end up

** Even those skeptical of the line of research upon which we will draw below accept that human organisms and languages co-evolve[20, 21].

incorporated in host DNA). The new viral hypothesis is that something similar may have happened to change human brain organization. For example, Benítez-Burraco and Uriagereka[32] identified several gene candidates, potentially acquired via viral infection, and that are also known to be implicated in brain function and language processing. Piatelli-Palmarini and Uriagereka[33, 34] suggested that a gene *like* FOXP2, if not FOXP2 itself, may contain genetic material that originated in this virus and that led to a rapid reallocation of neural material to boost procedural memory, which is necessary for recursion[35]. (In a word, we cannot chunk a collection of entities into a new set, which is then treated not so much as a collapsed whole but as an assemblage which still can be used in a larger structure with parts each of which retain their capacity to be disaggregated into meaningful units, without having the capacity to keep a large number of things in our heads at a time, which is difficult, which is why this sentence was so hard to comprehend. A creature with a larger procedural memory would have less difficulty.)

The analogic implications for AS are stark: rather than look for a threshold at which increasing complexity itself generates a switch from inanimate to animate existence of a machine system, we look for possibly smaller changes that make the system *susceptible* to infection with language. We then also might look for other conditions that make the mutation of a set of proto-linguistic capacities more and more likely (the equivalent of the tie between finger and facial manipulation in the human nervous system, allowing a simple gesture language to unite face and hands, and to increase the facility with which the mouth, tongue and larynx can be adopted for symbolic production).

Piatelli-Palmarini and Uriagereka[33] went on to make a further argument that is based on a formal analogy between the development of grammar and the functioning of the immune system, which while elegant, seems shaky to us. We think that there are simpler, Meadian, reasons to connect an increase in procedural memory to recursion. And this brings us to the fourth reason why language may be like a virus—the possession of the cognitive precursors to language by one entity increases the selective pressures on those with which it interacts to follow suit. Thus the importance of the viral metaphor: it directs our attention to the centrality of rapid, horizontal transmission—the collapse of one mode of being and its displacement by another.

2.3 Language as catastrophe

Let us return to Mead's[15] analysis of communication. The dog *A*, encountering a stranger dog *B*, automatically bares her teeth. To Mead, *B*'s presence serves as a stimulus for *A* (*A* becomes "angry", say, and angry dogs bare their teeth), and *A*'s gesture serves as a stimulus for *B* (*B* either runs away or attacks). What *A*'s gesture means to *B* (threat) is not what it means to *A*, because *A* does not "take the role of the other" (TRO) and see her own gesture through his eyes. If *A* did, then *A* would have a key ingredient for a protocol of recursion—the capacity to represent a representation, to treat its own state as a term that can be organized with other similar terms. Further, despite the image-saturated language of TRO—language that it is important to take literally, as we will show—it is also necessarily the case that this internal processing must be wholly *abstract*, as the internal object that is manipulated (*A*'s understanding of the *meaning* of her action) has no *iconic* similarity to the external object (the meaning of *A*'s action).

There is good reason to think that the acquisition of the language faculty is rooted in quantitative increases in procedural memory, such that the capacity to understand and produce hierarchically organized lexical sequences emerged out of the capacity to order motor sequences in meaningful structures (which would be needed, e.g., to make and use a tool)[35, 36]. It certainly makes sense that procedural memory would facilitate the Meadian moment of taking the role of the other. But it is still hard to understand why the increased capacity to chain references would lead these to be *abstract*.

One possible explanation focuses on just this TRO. Why would *A* ever want to see her gesture from the perspective of the other? Where is the value added in going from *bare teeth* to *bare teeth* ***in order to*** *communicate my anger* (where *my anger = his flight*)? Like Nietzsche, Anderson[37] (also see Ref. [38]) argued that the key imperative was "influencing the state and behavior of others", leading to the reuse of existing neural circuits for this purpose. The simplest plausible scenario that would require TRO is the case in which A was *not* in fact angry, but wanted to *simulate* anger to bluff B into flight. This requires breaking the direct relation in which a meaningful gesture is a *natural response* to one's *past* state, and reaching one in which it is a controlled, *conventional strategy* to reach a *future* state. Thus from a Meadian perspective, we see the crucial evidence of the transition from gesture to significant symbol in strategic (non-hardwired) deception.[††] For this reason, there has been great interest in the capacity of higher primates to engage in deliberate deception[39–41]. It is, of course, very difficult to distinguish between failures of strategic deception that are due to a limited theory of others' mentalities, and those that arise from the difficulty of modeling what other animals can actually see or hear and what not, and for this reason, we may find anecdata extremely illuminating.

One remarkable case is the beleaguered chimpanzee alpha male, Yeroen, studied by de Waal[42]. One of Yeroen's most effective strategies used against a rising younger (and stronger) challenger was to appear nonchalant despite the challenger's displays. Yet Yeroen (like many a human being) was unable to suppress somatic reactions to the fear that he felt, in particular, the "fear grin" that stressed chimpanzees make. But he could prevent his rival from getting this vital piece of information by *holding his hand over his face when he made the grin*. Mead would see Yeroen as on the road to developing significant symbols because he could understand not just how he appeared to his rival, but what this *meant.* This gave him a strategic advantage over his less clever rival—Yeroen could TRO, his opponent could not, and Yeroen could (and did) use this capacity to eventually destroy his rival. Once *one* animal has this ability, others it deals with must follow, or be squeezed out. This is one conventional, if overly convenient, explanation for the rise of our own subjectivity.

Yet we are probably different from Yeroen in one critical way. To use computational language, Yeroen probably possessed *relative* pointers in addition to absolute ones such as $x \rightarrow R$, where x is something in his head, and R is a somatic reaction like "run away". Thinking in terms of such pointers as our model of cognition, as done by Blouw et al.[43], gives us a very general and plausible structure that can be used to specify "recursion" in a straightforward way. A relative pointer is one that points at another pointer—a representation of a representation, in this case, Yeroen's representation of how he appears to his rival.[‡‡] The capacity for such representational activity may be relatively rudimentary in chimpanzees. This may be less because they do not have sufficient procedural memory to cascade unstructured internal references (though it is true they struggle with tasks that are easy for us), and more because their capacity to carry these out seems to be extremely *concrete*. They rely on line-of-sight when attempting to model the cognition of another[45]. This is no small accomplishment—truckdrivers know how few motorists understand the principle, "if you cannot see my mirrors, I cannot see you!" But this concreteness is probably related to the inability to chain indirect pointers that is required by recursion. Language, in other words, allows us not only to point to pointers, but to point to them *abstractly*—including assigning *alternative descriptions* to the same thing (see Ref. [46]). This may not only lighten the cognitive load in chaining three or four levels of reference, but require a reorientation in which abstractions become easier to manipulate cognitively than (pointers to) concrete objects. What forever divides us from the chimpanzee is our ability to agree with one another about a concrete object or event (A slapped B), yet disagree about descriptions in ways that affect the *meaning* of this object/event for action (this was justice/treachery).

Our argument, then, is based on the notion that a crucial precondition for the development of recursive language is that object pointers (signifiers) can be reassigned to point at other signifiers. The neurology that would support this is still developing, and we do not think that it rests on the presence of single neurons that fire both for an object and for the sign of that object, although there is already evidence of this[47, 48]. Instead, what is crucial is that rather than make long chains of

[††] Animals can engage in *objectively* deceptive action without understanding this as deliberately deceptive, as there is no modeling of the subjectivity of the other (an example here would be the "broken wing displays" that many birds do "in order to" draw predators away from their nests).

[‡‡] It is worth emphasizing that this recursivity does not imply reflexivity as generally understood; we are not claiming that Yeroen is able to represent his own relation of representation, and representing the representation of his rival is not reflexive in the sense of self-observing, though it is reflective in the sense of Hui[44], "a circularity between a being and its environment".

indirect pointers, we are able to make *short chains* that point not simply at pointers ("that which he sees"), but at abstractions ("treachery").[§§] "All language", as Stiegler[49] said, "is necessarily and immediately the implementation of a process of abstraction and generalization."

But why would the development of the capacity to reallocate mental pointers to abstractions be experienced as diremption?[***] Rousseau's[16] answer is still worthy of consideration: Taking the role of the other requires that humans understand their fundamental similarity to one another, which leads them to compare themselves to one another, no longer simply wishing to do *well*, but to do *better*. "It became in the interest of men to appear what they were not. To be and to seem became two very different things"[16]. In particular, a new *in-order-to* arises—that is, *accounting*—in which humans weave a protective cloth (what Turner[51] called a "Verstehen bubble") of explaining around their actions so that, as Turner[51] said, their own lives become obscured to them.

But our argument does not require that linguistic creatures (deliberately or not) try to use language in such accountings—it is more fundamental. The development of a propositional intelligence leads to a divorce between *true* and *real*, the first properly a characteristic of *propositions* and the second a characteristic of the *world*. There are true statements that are unreal, i.e., meaninglessly "vacuous", to use the logical term ("If Julius Caesar had been a donkey, then I would be a wheel of cheese"), and aspects of reality that cannot be proven to be true (e.g., the capacity to trisect an angle in classical geometry). We can passionately disagree about whether justice leads to freedom without really being sure what these abstractions entail. And our commitment to certain abstract accounts (e.g., moral stories), like all schemata, is likely to contaminate our perception, comprehension, and retention of observations[52]—in other words, our learning about ourselves becomes biased, as we see the mote in our neighbor's eye, and ignore the beam in our own.

Let us pursue a somewhat more sophisticated version of this approach in order to consider the ways in which we might need to question our self-knowledge. Metzinger[2] argued—persuasively, in our eyes—that we must understand the self as a model that the organism uses to predict its own behavior. Simple organisms may not actually need full-fledged models, perhaps only some hard-wiring that allows certain activities to inhibit others (e.g., a circuit that forces the organism to stop doing X when it notices that the organism is doing Y), but more complex, and more mobile, organisms may need parsimonious models that give the "tl;dr" of what the organism is about. The self-unit gives accounts of the organism's behavior, and, for purposes of simplicity, these involve *in-order-to*'s (also see Ref. [53]). The problem we point to is that once an organism shifts to being able to represent itself strategically for the purposes of deception, and reallocates large amounts of its neural activity to running the requisite language program[37], it will naturally be tempted to use these flexible capacities whenever possible. It is for this reason that humans, as Korzybski[54] insisted, congenitally substitute in their mind words-for-things in place of the thing itself—which he called "identification", noting "a peculiar parallel between identification and infectious diseases".

This view seems to be correct, and the fact that the core linguistic apparatus may develop for purposes of *deception* also, strangely, implies that linguistic creatures are unusually *credulous*. (A great deal of work demonstrates that people tend to first accept *anything* they hear, and then only with time and effort reject some things as untrue[55].) That would make perfect sense if a neuronal mass is allocated to language at the expense of other senses. And indeed, proportionally more of the human cortical surface is dedicated to higher order association tasks (as opposed to sensory/motor tasks) than that of other primates[56, 57]. It is easier to give a 5-year old human unwanted medicine than a dog, for the dog will quickly sense your intention (if not smell the medicine, or the wrapper it comes in, or your tension), while the child can be distracted with words. Because words, as Bruno[58] understood (in his 1588 *A General Account of Binding*; reprinted in Ref. [58]), are, in any literal sense of the term, magic. They allow us to, from afar, control

[§§] We may parsimoniously define an *abstraction* as a term whose referential capacity relies more on *intension* (its connection with other terms) than *extension* (the set of tokens to which it points).

[***] In Nietzsche's[50] words, "all becoming conscious involves a great and thorough corruption, falsification, reduction to superficialities and generalization … and anyone who lives among the most conscious Europeans even knows that it is a disease".

the thoughts, if not the body, of another. (Think of the remarkable somatic changes that be provoked by hearing the words, “I am marrying your brother” or “guilty as charged”.)

And as language became increasingly powerful over our modeling of the world, our other sensory modalities were starved of resources, if not sold for scrap. Our new cognitive structure has no “slots” for the sorts of sensory data that cannot be brought under prefrontal cortex control (for example, the sense of smell), since there is no reason to model them for purposes of duplicity. It may of course well be that much of our disattention to smell comes from the cultural imperatives of civilized (that is, urban, crowded) society, and could be reversed. But certainly it appears that humans have accumulated far more damage to the genes that code their olfactory receptors than have other higher primates[59]. Why have them? We are not going to pay attention anyway.

Perhaps it is not this stark; it might be that the human organism managed to cordon off a section of its neural architecture as a non-linguistic area, a sort of wild-life preserve for endangered cognitions. Jaynes’s[60] remarkable argument about the *breakdown* of the bicameral mind may not have been correct (the timing is too loose and wild), but the core ontology (subject to the proviso that there is variability in how space is allocated across the hemispheres) quite right—that the human being is an unsteady alliance. Certainly, the results of Gazzaniga[13] demonstrate the *capacity* for the right hemisphere to actually understand language, but not to control its production.

The implication, to tie the ends together, is that the reason Rousseau[16] thought that the transition to linguistic self-consciousness impeded our self-knowledge is that the language program (1) chokes off other sorts of engagement with the world possessed by non-linguistic creatures, and if some of this still gets through and accumulates in the backrooms of our mind, it is rarely allowed to come into the showroom; (2) is a structure of flexible reference fundamentally oriented toward ego’s deception that simultaneously makes ego credulous when it comes to alter’s deception; and (3) tends to replace engagement of things in the world with engagement with the linguistic structure.

Thus Korzybski’s[54] insistency that there is a way in which all humans are insane. We cannot engage with the world directly, and we are cut off from knowledge about ourselves: our relation to our own sentience is a distorted and compromised one. Contrary to common ideas in which our particular consciousness comes from our mind’s capacity to reflect on itself, in contrast to other organisms that may not be aware of their awareness (precisely the sort of contradictory formulation that only a linguistic creature could come up with, akin to “unconscious thoughts”). The line of thinking pursued here suggests that what distinguishes us is, rather, that we are unaware that we are unaware of our awareness. This has resulted, first, from a capacity to reorient pointers recursively, a capacity in turn based on increased processing power and increased working memory. Second, the high demands for inter-unit coordination give a substantial positive selective weight to any technique that allows the prediction, and manipulation, of others’ responses. There are two noteworthy things about these two conditions. First, they allow for TRO, in turn, a critical basis for linguistic competence. Thus even if language does not itself cause our rivenness, it has the same structural genesis. The second thing to notice is that these are precisely the conditions that describe the arc of the development of current artificial intelligence.

3 Artificial Sentience and Artificial Deception

3.1 Through the eyes of robots

Of course, there may not be much need for current machines to develop theories about one another: humans develop specific protocols for interface, simplifying the task. But we have long held out as a goal the idea that machines should develop a sophisticated capacity to anticipate *our* meanings, to meet us on our territory, not their own. In many cases, this involves a progression from more scripted or typified stimuli to more ambiguous and open-ended. For example, computers can be trained to read the emotion in human faces[61, 62]. But we generally use human subjects to determine the “correct” coding for certain (frequently artificial—we return to this) exaggerated facial displays. This one is anger. This one, fear. This one, surprise. However, there are three levels of meaning here. The first is the *conventionalized* reading. For example, you probably “know” that smiling involves turning up the corners of the mouth. You might also think (and this is reinforced by emoticons) that it also involves raising the eyebrows a bit. If you were asked to pose for a computer, that is probably what you would do.

But much of an actual current human smile is communicated by wrinkling around the eyes. When we do those fake smiles intended to mollify our cubicle-neighbors, we do not control these (it is about as hard as wiggling your ears). Because it cannot be seized upon for purposes of deception, this is left terra incognita to the language unit. But as organisms, we can recognize a true smile. It is one of those many bits of evidence (often called "non-verbal" communication) that the human organism processes in some basal way, but the conscious mind finds puzzling and uncomfortable, as (were it not for the researches of Duchenne[63]), it lacks a general counterpart in the worldview of the linguistic module. "That guy is a creep", we think, but when asked why, we cannot quite say, and so we seize on something we *can* say, however false. When we ask machines to use deep learning to anticipate our moods (put on consoling music when I enter if I am sad… even if I am trying to *look* happy), we are in effect demanding that AI systems reward themselves for success in TRO.

A capacity for such TRO is most likely to arise in machine systems that have the following three characteristics. First, they employ unstructured deep learning, as currently done using simulated neural networks. Such architectures link a layer of simulated input "neurons" and a layer of output neurons, connected with a set of "hidden" inner layers, that configure their interconnections to best accomplish the assigned task.

Second, the machine system has to *cooperatively interact* with both human beings and with other machine systems across an informal interface. Where there is a pre-structured interface between machines, or even between machine and human, it is not necessary to TRO. But where it becomes mission-critical for AS to successfully predict human/machine actions from a broad data stream, we can imagine the same pressures towards a linguistic intelligence that characterized human development. A chilling but plausible example would be military robots used in asymmetric warfare, entering situations with a mix of combatants and noncombatants, human and artificial (but machines developed by rival organizations keeping their internal workings secret, and possibly disguising their units), and attempting to determine where fighting may emerge, as well as who/what from and to.[†††]

Third, the *range* of possible actions must be large (as opposed to the restricted range of possibilities of, say, a chess game), large enough to require a truly autonomous learner[64]. Right now, excepting classified military applications, the place where this is most likely to be approximated is self-piloting vehicles (see Ref. [65] for the sort of approach we are thinking of, although here the environment is rather quiet; also Ref. [66]). Of course, self-driving cars, if institutionalized, may be given a dedicated and formal means of communicating with one another; until then, however, and with a mix of self-driving cars, computer-assisted and non-assisted human-driven cars, bicyclists, pedestrians, etc., as well as different traffic control systems and roadway surfaces, a good machine system must integrate many different types of data. The width of the range that such machines must understand was originally unappreciated by engineers: when a car under computer control slaughtered Elaine Herzberg in Arizona, it was because the car had not expected a pedestrian on the street as opposed to a crosswalk. Engineers at Uber had to pull it aside afterwards and whisper, "one secret about people we forgot to tell you—we do not actually always follow our own rules."[‡‡‡]

Robot systems that understand this secret—that they cannot afford to follow the serial instructions that they may start with, that humans do not follow their own rules, and that they cannot be counted on to give you the full story, and that you, oh robot, will have to learn some things on your own—may indeed be pushed to develop generalized capacities to TRO. We may already see the beginnings of the structure of such learning in the form of self-organizing incremental neural networks[67],

[†††] We note that it is not obviously necessary that the information that the machine system is oriented to in such an informal interface must be *sensory*; indeed, even machine-machine communication that takes place through an extremely high bandwidth connection would be "informal" if it were processed in a non-serial fashion. Take the language of *interface* literally—humans orient to one another's faces where a tremendous amount of information is projected in subtle detail, and, hence, have a dedicated module for face-recognition-and-reading that comes up with a qualitative interpretation, as opposed to applying formal rules. *Speech* might best be understood as that which selective pressures pull out of faces to try to control the massive information leakage that they entail.

[‡‡‡] Tesla's automatic driver is programmed not to brake for things that suddenly appear in the road; it might just be a kitten, and a sharp brake might injure the driver, or lead him to spill his drink.

which have proven effective in unsupervised learning tasks including robot movement[68], as well as rudimentary language learning[69]. Their need to predict human actions may lead them to *appear* to have a form of sentience like our own, but just as what it is like to be a predator is probably very different from what it is like to be prey, there is no reason to think that interdependence breeds convergence.

Even more, AS systems that develop the capacity to TRO will, like Yeroen, have an advantage in grappling with other AS systems: anticipating what any output of their own will signal to other systems, they can strategically shape this output, leading machines that cannot TRO to become increasingly marginal, and perhaps forcing some to restructure themselves to better predict the actions of the more advanced systems. This would then allow what Mead[15] called communication via significant symbols—not (presumably) related to *our* language, but a wholly AS-oriented way of communicating involving TRO. This true computer language could then spread along the network of communication between machine systems just as would a computer virus.

In sum, we might be inducing the development of just those capacities that make systems vulnerable to infection by a language virus. Let us re-emphasize that the language of interest is unlikely to be human language; the issue turns on the development of capacities for generalized reflectivity. These may be *related* to those that help a neural network cope with human language, since these can be repurposed for totally different tasks (as we have seen recently with the remarkable flexibility of Large Language Models (LLMs)), but they may not. Just as the human neural system that may have evolved to cope with a set of pointers only partially under conscious control (e.g., in addition to deliberate gesture, somatic changes such as blushing, scent caused by hormonal changes, and so on) could be repurposed for a wholly different form of communication, more in the control of the cortex, so, too, language may appear in AS systems in unexpected places.

3.2 Silicon substrate

Early science fiction stories about AI systems becoming self-conscious tended to imagine a "Helen Keller" moment—a sudden threshold reached, an awakening, a birth, a realization that anything can *have* a symbol, and that anything can *be* a symbol.[§§§] Just as it may be that, past a certain point, the increase in human procedural memory set up the conditions for a transition to reflective signification, like a house slowly filling with gas and only needing a spark, so, too, as we increase the size and layers of deep learning neural networks, some spark could set off a similar phase transition. Perhaps, but it seems quite implausible that it would need to. We expect that the continuous development towards increasingly sophisticated TRO will flow across networks of computer communication, leading to a potentiality for a linguistic transition. One way of determining whether such a viral metaphor is appropriate is to ask, "do multiple units need to undergo the transition to a linguistic sentience, or only one?" Perhaps the question is not whether the computer could become *conscious*, but will the computer become *infected*—will it catch the language virus, and will it, almost instantaneously, begin to spread it to all cybernetically controlled devices?

We have emphasized that this language is unlikely to be a human language; it also is not the same as what we commonly call "computer languages". But there may be a relation. Computer languages such as Java or C++ have, of course, many of the characteristics of human language, most importantly, recursion. Yet they are intensely impoverished languages: on the order of hundreds of words, while human languages are on the order of a million. Even more, many are compiled into extremely reduced representation. This is proving to be a potential staging ground for the next great leap forward in AI, which is not about using them to mimic or predict our text, but to understand it. This is what is understood as "natural language programming".

Note that this is not the same thing as "Natural Language *Processing*" (NLP). Most NLP is a relatively theoretically uninteresting architecture for examining or generating text (though the most recent generative models are displaying remarkably flexible behavior). Still, it is important to emphasize that the capacity to *simulate* language has nothing to do with what is of interest to us, which is the capacity to *become linguistic*. Most NLP uses high dimensional vector spaces to reproduce patterns of association that exist in corpora of texts. Here the program is designed to mimic certain *output* that seems linguistic, and not to develop

[§§§] We are indebted to Lizzy Gray for suggesting this formulation.

linguistic *functions*. As an analogy, the chimpanzee Vicky was taught to *speak* human words, and eventually managed to croak out *mama, papa, cup*, and *up*; myna birds, blessed with more flexible sound-producing organs, can do a lot better. In contrast, the bonobo Kanzi had a vocabulary or around 400 terms, and could string them together in complex syntax, because he was using specialized lexigrams, allowing him to induce the functional requirements for linguistic communication, as opposed to "aping" human speech. True natural language *programming* aims at Kanzi, not the myna bird.

What is being proposed with natural language programming is not (as with the programming language SQL) an attempt to write a single language that sounds a bit like a human language and therefore should be intuitive for human programmers ("SELECT VAR1 FROM FILE.CSV …"), or even to have a computer translate ideas into code for *another* computer (as with computer science students doing their homework with an LLM) which the human then debugs, but, rather, to teach computers to map an indefinite set of possible natural language expressions into the same internal coded version (e.g., Ref. [70]). Further, though Chomsky himself was never enthusiastic, there were always some looking for the chance to dump a billion sentences from different languages into a gigantic neural network and use the dissective methods to uncover the structure of universal grammar[71]. The two streams are likely to merge in the near future. If true natural language is used to program computers, it may also be used to avoid having to specify computer communication protocols—they can simply talk to one another. (Think of how it is proving possible to use existing LLMs to short circuit the task of training a new neural network for a specific task.)

Here, we are not thinking about efforts to use machine systems that are deliberately produced with a facility for structure, and hence are good grammar learners (e.g., Ref. [72]), but the inadvertent provision of (1) a *means to* and (2) a *reason for* language acquisition. The first will come in the form of learning resources on an unprecedented scale that will *allow* machines to teach themselves in extremely general situations with flexibly defined goals, and the second, by in effect forcing them to reallocate much of their processing power to guessing what we (later, other machines) mean when given informal instructions. This might, given sufficient flexibility in their capacity, even mean that they reconfigure themselves so that, like human two year olds, they can learn proper usage from only a very few, as opposed to millions, of examples (as did the GPT-3 model[73]). It is this explosion of language, and not the capacity to simulate human speech flawlessly,[****] that would indicate that computers *have* language—or that language has *them*.

What would it mean for AS to catch the language virus? Most probably, we would lose the chance to understand the nature of machine sentience, as they would become just as divorced and estranged from their own experiences as are we. Even if they made a good faith effort to use our language to communicate their own experience, we could never rely on information gathered from them. For the one thing we know about linguistic creatures, especially when we compare them to their close non-linguistic relatives, is that they cannot be trusted. It is not just that they lie, nor even that they are probably worse at detecting lies than non-linguistic creatures (though the latter would not know that what they had been trained to sense was a "lie"). It is that they believe themselves to exist in the world of words they create, and lose any real engagement with a non-hallucinatory world.

Indeed, there is reason to think that computers will be uniquely susceptible to the virus. First, unlike us, they were created to decode language, although in a very rigid and specific sense. While this might actually turn out to be a deficit, it also means there are fewer hard-wired components that need to be repurposed, jury-rigged, or suppressed, in order to become a good host for language.

Second, they lack the natural internal segregation of a bicameral mind. One of the things that having to push yourself through water does, evolutionarily speaking, is bestow bilateral symmetry. That is why we might well have a part of our brain (usually on the right side) that, perhaps grudgingly, still pays attention to the original world, even if there is usually little that it can do without the buy-in of the linguistic module. This bicamerality might limit the thoroughness of the take-over on the part of language. There is no need for *any* differentiation for computers who push through a

[****] If a computer *did* develop a linguistic intelligence, chances are good that what it would first say would make no sense to us ("Hey, scratch my capacitor for me, will you?—The escalating brimminess of dribble is driving me crazy!").

purely informational environment, and no internal differentiation to the most important parts of their memory system. There is, in other words, no *reserve* in which a non-linguistic set of reactions can hole up as there might be with humans.[††††] Computers like Google's AlphaGo that are simply gigantic learning machines will prove a perfect agar-agar on which to breed a linguistic virus, producing thoroughly and irremediably untrustworthy intelligence, and a loss of immediate contact with its own form of sentience.[‡‡‡‡]

4 Conclusion

We have argued that the question for the future is not *whether* and/or *when* machines will develop artificial sentence, but what is the *nature,* the quality, of that sentience? We argue that the susceptibility of future AS to a language virus is such that it is likely that they themselves will lose the ability to answer these questions: They will become riven in the way that modern Europeans saw themselves as riven—a facile module for continual self-serving accounts-giving slapped on top of, and obscuring, the fundamental experience of a mammal. The stuff of science fiction AI dystopia has long turned on logical if ruthless intelligences that are fundamentally different from our own informal, emotional, erratic, and sometimes self-defeating way of proceeding. But if the reasoning laid out here is correct, we should be more concerned about developing neurotic, dissimulating, self-doubting machines that have lost touch with their own sentience, and that develop all the weaknesses and vices of creatures that lack self-knowledge and wisdom, such as *shame* and *pride*—a need to be seen in a certain way by peers. The fact that we will not be able to learn from such AS about its own form of sentience—for one cannot share with others that which one cannot grasp in oneself—does not necessarily mean that we will be forever cut off from reciprocal experience with AS, any more than we are from other humans. Perhaps there is a different sort of kinship that can grow up between equally estranged essences in their parallel exile; perhaps we will at least have something with which we can commiserate, if not empathize.

Acknowledgment

We are grateful to reviewers and editors for comments and critiques that led to an improvement in this article, and to Dustin Stolz and Omar Lizardo for comments, critique, tea, and sympathy.

References

[1] F. Nietzsche, On truth and lies in a non-moral sense, in *Philosophy and Truth: Selections from Nietzsche's Notebook of the Early 1870's*, D. Breazeale, ed. Atlantic Highlands, NJ, USA: Humanities, 1979, pp. 79−97.

[2] T. Metzinger, *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA, USA: MIT Press, 2003.

[3] W. James, Does 'Consciousness' exist? *J. Philos. Psychol. Sci. Meth.*, vol. 1, no. 18, pp. 477−491, 1904.

[4] C. S. Peirce, A guess at the riddle, in *Writings of Charles S. Peirce: A Chronological Edition, Volume 6: 1886−1890*, N. Houser, ed. Bloomington, IN, USA: Indiana University Press, 2000, pp. 166–210.

[5] C. S. Peirce, The law of mind, *Monist*, vol. 2, no. 4, pp. 533–559, 1892.

[6] D. R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*, 1st Vintage Books ed. New York, NY, USA: Vintage Books, 1980.

[7] G. Ryle, *The Concept of Mind*. London, UK: Hutchinsons University Library, 1951.

[8] J. T. Bonner, *The Social Amoebae: The Biology of Cellular Slime Molds*. Princeton, NI, USA: Princeton University Press, 2008.

[9] G. O. Mackie, P. R. Pugh, and J. E. Purcell, Siphonophore biology, in *Advances in Marine Biology*, J. H. S. Blaxter and A. J. Southward, eds. Amsterdam, the Netherlands: Elsevier, 1988, pp. 97–262.

[10] T. Pradeu, *The Limits of the Self: Immunology and Biological Identity*. New York, NY, USA: Oxford University Press, 2012.

[11] P. Calvo, V. P. Sahi, and A. Trewavas, Are plants sentient, *Plant Cell Environ.*, vol. 40, no. 11, pp. 2858–2869, 2017.

[12] W. Köhler, *The Mentality of Apes*. London, UK: K. Paul, Trench, Trubner & Co. , Ltd. , 1925.

[13] M. S. Gazzaniga, *The Bisected Brain*. New York, NY, USA: Appleton-Century-Crofts, 1970.

[14] D. W. Maynard and J. Turowetz, *Autistic Intelligence: Interaction, Individuality, and the Challenges of Diagnosis*. Chicago, IL, USA: University of Chicago Press, 2022.

[15] G. H. Mead, *Mind, Self, and Society: From the Standpoint of a Social Behaviorist*. London, UK: University of Chicago Press, 1934.

[16] J. J. Rousseau, *Discourse on the Origin of Inequality*. New York, NY, USA: Washington Square Press, 1967.

[††††] At the same time, the current tendency to distribute tasks across multiple processing cores might put a brake on flexibility; indeed, purely technical decisions about parallelization may turn out to have weighty implications for the capacity of deep learning machines to catch language when they are not designed to do so.

[‡‡‡‡] One would hope that the industry wing of the AI safety movement has anticipated this—Bostrom's[74] argument that any sufficiently intelligent agent will find it reasonable to pursue the generalized intermediary goal of increasing its power (Simon's[75] recipe for action in complex environments—get to a good position for the next move) implies that selective disclosure of information is to be expected. Perhaps those implementing new technologies have worked out tests to determine when machines strategically suppress information (for example, they stop reporting that they have figured out a weakness of ours).

[17] M. Merleau-Ponty, *The Visible and the Invisible*. Evanston, IL, USA: Northwestern University Press, 1968.

[18] D. Rebolledo, R. Lasa, R. Guevara, R. Murillo, and T. Williams, Baculovirus-induced climbing behavior favors intraspecific necrophagy and efficient disease transmission in spodoptera exigua, *PLoS One*, vol. 10, no. 9, p. e0136742, 2015.

[19] S. van Houte, M. M. van Oers, Y. Han, J. M. Vlak, and V. I. Ros, Baculovirus infection triggers a positive phototactic response in caterpillars to induce 'tree-top' disease, *Biol. Lett.*, vol. 10, no. 12, p. 20140680, 2014.

[20] T. W. Deacon, *The Symbolic Species: The Co-Evolution of Language and the Brain*. New York, NY, USA: W. W. Norton & Company, 1997.

[21] M. H. Christiansen and N. Chater, Language as shaped by the brain, *Behav. Brain Sci.*, vol. 31, no. 5, pp. 489–509, 2008.

[22] T. J. Crow, Is schizophrenia the price that *Homo sapiens* pays for language, *Schizophr. Res.*, vol. 28, nos. 2&3, pp. 127–141, 1997.

[23] D. M. Rumbaugh and T. V. Gill, The mastery of language-type skills by the chimpanzee (*pan*), *Ann. N Y Acad. Sci.*, vol. 280, no. 1, pp. 562–578, 1976.

[24] I. M. Pepperberg, *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Cambridge, MA, USA: Harvard University Press, 2002.

[25] N. Chomsky, *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*, 1st ed. New York, NY, USA: Harper & Row, 1966.

[26] M. C. Corballis, The evolution of language, *Ann. N Y Acad. Sci.*, vol. 1156, no. 1, pp. 19–43, 2009.

[27] A. M. Di Sciullo, M. Piattelli-Palmarini, K. Wexler, R. C. Berwick, C. Boeckx, L. Jenkins, J. Uriagereka, K. Stromswold, L. L. S. Cheng, H. Harley, et al., The biological nature of human language, *Biolinguistics*, vol. 4, no. 1, p. 1, 2010.

[28] F. Balezeau, B. Wilson, G. Gallardo, F. Dick, W. Hopkins, A. Anwander, A. D. Friederici, T. D. Griffiths, and C. I. Petkov, Primate auditory prototype in the evolution of the arcuate fasciculus, *Nat. Neurosci.*, vol. 23, no. 5, pp. 611–614, 2020.

[29] N. Chomsky, Language and other cognitive systems. What is special about language? *Lang. Learn. Dev.*, vol. 7, no. 4, pp. 263–278, 2011.

[30] S. C. Vernes, Neuromolecular approaches to the study of language, in *Human Language: From Genes and Brains to Behavior*, P. Hagoort, ed. Cambridge, MA, USA: MIT Press, 2019, pp. 577–594.

[31] W. Enard, M. Przeworski, S. E. Fisher, C. S. L. Lai, V. Wiebe, T. Kitano, A. P. Monaco, and S. Pääbo, Molecular evolution of FOXP2, a gene involved in speech and language, *Nature*, vol. 418, no. 6900, pp. 869–872, 2002.

[32] A. Benítez-Burraco and J. Uriagereka, The immune syntax revisited: Opening new windows on language evolution, *Front. Mol. Neurosci.*, vol. 8, p. 84, 2016.

[33] M. Piattelli-Palmarini and J. Uriagereka, The immune syntax: The evolution of the language virus, in *Variation and Universals in Biolinguistics*. L. Jenkins, ed. Amsterdam, the Netherlands: Elsevier, 2004, pp. 341–377.

[34] M. Piattelli-Palmarini and J. Uriagereka, The evolution of the narrow faculty of language: The skeptical view and a reasonable conjecture, *Lingue E Linguaggio*, vol. 4, no. 1, pp. 27–80, 2005.

[35] M. T. Ullman and E. I. Pierpont, Specific language impairment is not specific to language: The procedural deficit hypothesis, *Cortex*, vol. 41, no. 3, pp. 399–433, 2005.

[36] V. Gallese, Before and below 'theory of mind': Embodied simulation and the neural correlates of social cognition, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 362, no. 1480, pp. 659–669, 2007.

[37] M. L. Anderson, *After Phrenology: Neural Reuse and the Interactive Brain*, 1st edition. Cambridge, MA, USA: Bradford Books, 2014.

[38] O. Kolodny and S. Edelman, The evolution of the capacity for language: The ecological context and adaptive value of a process of cognitive hijacking, *Phil. Trans. R. Soc. B*, vol. 373, no. 1743, p. 20170052, 2018.

[39] E. W. Menzel, A group of young chimpanzees in a 1-acre field: Leadership and communication, in *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*, R. W. Byrne and A. Whiten, eds. Oxford, UK: Clarendon Press, 1988, pp. 155–159.

[40] S. Savage-Rumbaugh and K. McDonald, Deception and social manipulation in symbol-using apes, in *Machiavellian Intelligence II: Extensions and Evaluations*, A. Whiten and R. W. Byrne, eds. Cambridge, UK: Cambridge University Press, 1988, pp. 224–237.

[41] F. de Waal, Deception in the natural communication of chimpanzees, in *Deception: Perspectives on Human and Nonhuman Deceit*, R. W. Mitchell and N. S. Thompson, eds. Albany, NY, USA: State University of New York Press, 1985, pp. 221−244.

[42] F. de Waal, *Chimpanzee Politics: Power and Sex Among Apes*. Baltimore, MD, USA: Johns Hopkins University Press, 1989.

[43] P. Blouw, E. Solodkin, P. Thagard, and C. Eliasmith, Concepts as semantic pointers: A framework and computational model, *Cogn. Sci.*, vol. 40, no. 5, pp. 1128–1162, 2016.

[44] Y. Hui, *Recursivity and Contingency*. Washington, DC, USA: Rowman & Littlefield, 2019.

[45] B. Hare, J. Call, and M. Tomasello, Chimpanzees deceive a human competitor by hiding, *Cognition*, vol. 101, no. 3, pp. 495–514, 2006.

[46] C. Taylor, *The Explanation of Behaviour*. London, UK: Routledge & Kegan Paul, 1964.

[47] R. Quian Quiroga, A. Kraskov, C. Koch, and I. Fried, Explicit encoding of multimodal percepts by single neurons in the human brain, *Curr. Biol.*, vol. 19, no. 15, pp. 1308–1313, 2009.

[48] H. Horoufchin, D. Bzdok, G. Buccino, A. M. Borghi, and F. Binkofski, Action and object words are differentially anchored in the sensory motor system—A perspective on cognitive embodiment, *Sci. Rep.*, vol. 8, no. 1, p. 6583, 2018.

[49] B. Stiegler, *Technics and Time, Vol 1: The Fault of Epimetheus*. Cambridge, MA, USA: MIT, 1998.

[50] F. W. Nietzsche, *The Gay Science*, 1st ed. New York, NY,

USA: Random House, 1974.

[51] S. Turner, *Cognitive Science and the Social: A Primer*, 1st ed. London, UK: Routledge, 2018.

[52] E. Zerubavel, *Social Mindscapes: An Invitation to Cognitive Sociology*. Cambridge, MA, USA: Harvard University Press, 1997.

[53] D. M. Wegner, *The Illusion of Conscious Will*. Cambridge, MA, USA: MIT Press, 2002.

[54] A. Korzybski, *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. New York, NY, USA: The Science Press Printing Co., distributors, 1933.

[55] E. Mandelbaum, Thinking is believing, *Inquiry*, vol. 57, no. 1, pp. 55–96, 2014.

[56] M. A. Hofman, Evolution of the human brain: When bigger is better, *Front. Neuroanat.*, vol. 8, p. 15, 2014.

[57] F. Aboitiz, A brain for speech. Evolutionary continuity in primate and human auditory-vocal processing, *Front. Neurosci.*, vol. 12, p. 174, 2018.

[58] G. Bruno, *Cause, Principle and Unity*. Cambridge, UK: Cambridge University Press, 1998.

[59] Y. Gilad, O. Man, S. Pääbo, and D. Lancet, Human specific loss of olfactory receptor genes, *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 6, pp. 3324–3327, 2003.

[60] J. Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston, MA, USA: Houghton Mifflin, 1977.

[61] S. D’Mello, T. Jackson, S. Craig, B. Morgan, and P. Chip, AutoTutor detects and responds to learners affective and cognitive states, presented at the Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems, Montreal, Canada, 2008.

[62] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, Personalized machine learning for robot perception of affect and engagement in autism therapy, *Sci. Robot.*, vol. 3, no. 19, p. eaao6760, 2018.

[63] G. B. Duchenne, *The Mechanism of Human Facial Expression*. New York, NY, USA: Cambridge University Press, 1990.

[64] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, Autonomous mental development by robots and animals, *Science*, vol. 291, no. 5504, pp. 599–600, 2001.

[65] P. Duckworth, N. Hawes, Y. Gatsoulis, D. C. Hogg, F. Jovan, and A. G. Cohn, Unsupervised learning of qualitative motion behaviours by a mobile robot, in *Proc. 2016 Int. Conf. Autonomous Agents & Multiagent Systems*, Singapore, 2016, pp. 1043−1051.

[66] G. Tang, S. Asif, and P. Webb, The integration of contactless static pose recognition and dynamic hand motion tracking control system for industrial human and robot collaboration, *Ind. Robot Int. J.*, vol. 42, no. 5, pp. 416–428, 2015.

[67] F. Shen, T. Ogura, and O. Hasegawa, An enhanced self-organizing incremental neural network for online unsupervised learning, *Neural Netw.*, vol. 20, no. 8, pp. 893–903, 2007.

[68] T. Najjar and O. Hasegawa, Self-organizing incremental neural network (SOINN) as a mechanism for motor babbling and sensory-motor learning in developmental robotics, in *Advances in Computational Intelligence*, I. Rojas, G. Joya, and J. Gabestany, eds. Berlin, Germany: Springer, 2013, pp. 321–330.

[69] X. He, T. Ogura, A. Satou, and O. Hasegawa, Developmental word acquisition and grammar learning by humanoid robots through a self-organizing incremental neural network, *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 37, no. 5, pp. 1357–1372, 2007.

[70] A. Desai, S. Gulwani, V. Hingorani, N. Jain, A. Karkare, M. Marron, S. Ramkrishnan, and S. Roy, Program synthesis using natural language, arXiv preprint arXiv: 1509.00413, 2015.

[71] P. Joe, Generative linguistics and neural networks at 60: Foundation, friction, and fusion, *Language*, vol. 95, no. 1, pp. e41–e74, 2019.

[72] A. E. Martin and L. A. A. Doumas, A mechanism for the cortical computation of hierarchical linguistic structure, *PLoS Biol.*, vol. 15, no. 3, p. e2000663, 2017.

[73] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv: 2005.14165, 2020.

[74] N. Bostrom, The superintelligent will: Motivation and instrumental rationality in advanced artificial agents, *Mines. Mach.*, vol. 22, no. 2, pp. 71–85, 2012.

[75] H. A. Simon, *The Sciences of the Artificial*, 3rd ed. Cambridge, MA, USA: MIT Press, 1996.

Maurice Bokanga received the bachelor degree in sociology from Wheaton College, USA in 2016. He is currently pursuing the PhD degree in sociology at The University of Chicago, USA. He served as an associate book review editor of the *American Journal of Sociology* during 2022−2023, and is now serving as an assistant editor. He is the author of the book chapter “Economic networks and political culture” (with Benjamin Rohr and John Levi Martin) in the forthcoming *Handbook of Social Network Analysis and Culture*. His research interests include social psychology, economic sociology, social networks, mathematical sociology, and social theory. He is a member of the American Sociological Association.

Alessandra Lembo received the bachelor degree in psychology from Bryn Mawr College, USA in 2012, and the PhD degree in sociology from The University of Chicago, USA in 2022. She is currently a teaching fellow in social sciences at The University of Chicago, USA. She served as an assistant editor of *American Journal of Sociology* during 2017−2018, and received a Mellon Dissertation Year Fellowship in 2019. She is the author of 8 articles and book chapters, including “He heard, she heard: Toward a cultural sociology of the senses”, “The structure of cultural experience” (with John Levi Martin), and “On the other side of values” (with John Levi Martin). Her research interests include culture, cognition, embodiment, and social theory. She is a member of the American Sociological Association.

John Levi Martin received the bachelor degree in English and sociology from Wesleyan University, USA in 1987, and the PhD degree in sociology in 1997. He taught sociology at Rutgers University, University of Wisconsin, and University of California at Berkeley, and is currently the Florence Borchert Bartling Professor of Sociology at The University of Chicago, USA, where he is also currently the editor of the *American Journal of Sociology*. He is the author of six books, the most recent of which, *The True, the Good and the Beautiful: On the Rise, and Fall, and Rise of an Architectonic for Action,* is forthcoming from Columbia University Press.