Biosemiotics (2024) 17:655–670
https://doi.org/10.1007/s12304-024-09565-y

RESEARCH

Res Cogitans – The Evolution of Thinking

Patrik Lindenfors[1,2,3]

Received: 4 May 2023 / Accepted: 16 April 2024 / Published online: 29 May 2024

Abstract

A somewhat prominent view in the literature is that language provides opportunity to program the brain with 'cognitive gadgets', or 'virtual machines'. Here, I explore the possibility that thinking itself – internal symbolic responses to stimuli that are either intrinsic or extrinsic, and computational procedures that operate on these internal symbolic representations – is such a software product rather than just an emergent phenomenon of the brain's hardware being 'complex enough', or the brain processing information in a manner that is 'integrated enough'. I also present a testable hypothesis that would indicate the presence of such a thought-gadget, and briefly overview some evolutionary pre-requisites for its existence. Further, I explore some consequences the existence of such a gadget would entail for our understanding of consciousness. The nature of the gadget is left unspecified as the article is not a blueprint for the thinking gadget, but an argument in favor of its existence.

Keywords Thinking · Consciousness · Evolution · Cultural evolution · Cognitive gadgets · Memetics · Language

✉ Patrik Lindenfors
patrik.lindenfors@iffs.se

[1] Institute for Futures Studies, Box 591, Stockholm SE-101 31, Sweden

[2] Centre for Cultural Evolution, Department of Psychology, Stockholm University, Stockholm SE-106 91, Sweden

[3] Department of Zoology, Stockholm University, Stockholm SE-106 91, Sweden

Premises and Qualifications

It has been proposed by several workers that thinking, or even consciousness, is the product of information processing on a higher abstraction level combined with downwards causality, i.e., mental models or representational structures in the mind and computational methods that act on those models or structures. For example, Daniel Dennett has stated that 'Human consciousness is itself a huge complex of memes…' (Dennett, 1991, p. 210) and comparable propositions have been explored by e.g., Fodor (1975), Hofstadter (1979), Putnam (1980), Block (1995), Blackmore (2003), and Heyes (2018), to name but a few. Below, I point to some consequences of such a viewpoint.

First, however, any discussion about consciousness immediately runs into definition-problems. What, exactly, are we talking about? There are a variety of distinct but related meanings of the concept of consciousness, such as for example being *aware of and thus reflect on qualitative states* ('qualia'), which makes consciousness that which makes you *experience*, for example, the redness in red instead of only registering a wavelength. Consciousness can also be about phenomenological states or something that goes under the notion of 'narrative consciousness', that is, awareness streams. Other related definitions focus on the question if there is *something it is like it is to be* an animal, a computer, or a human being – subjective experiences, how the world *seems to me* (Blackmore & Troscianko, 2018; van Gulick 2022). It has also been emphasized that there are hard and easy aspects of understanding consciousness, where 'the hard problem of consciousness' entails understanding how subjective experiences arises from material processes (Chalmers, 1995).

Problems of understanding consciousness arise partly because there are currently no externally detectable indicators that can unequivocally demonstrate who (or what) is conscious and who (or what) is not. The only thing one can be reasonably sure of is one's own consciousness. It is for the same reasons currently impossible to determine if animals are conscious and if so, which ones. Animals undoubtedly cycle through periods being both awake and asleep, something that in people usually distinguishes states of consciousness and non-consciousness. But patients with brain injuries can also go through cycles of being awake and being asleep without anyone seemingly being 'at home'.

In what follows, I therefore mainly, but not exclusively, limit the discussion to *thinking*, here defined as internal symbolic responses to symbolic stimuli that are either intrinsic or extrinsic, and computational procedures that operate on these internal symbolic representations. I will return to the question of consciousness at the end of the paper. Below follow some premises for the rest of the article.

- Thinking is not just any information processing, but a *certain kind* of information processing. Otherwise, thermostats would carry out minimal thought processes, and we have reasons to believe that they do not, as they are completely obedient to their settings, carry out no processes in the absence of external stimuli, and show no signs of temporal adaptive change of internal states (learning). Further, most information processing that takes place in the brain does not result in thinking.

- If thinking is a certain kind of information processing, it is an ability that seems to be the *gained during childhood*, probably *a product of learning* (Heyes 2018). It seems we are not capable of thought from birth, as we do not form autobiographical memories until several years after birth and would not be able to pass the Turing test as babies. Sensory pain and pleasure responses are recognizably similar in small children as in adults, as are many responses to visual, auditory, gustatory, or olfactory stimuli. Thus, raw experiences seem not to be sufficient for thought. Animals also have recognizably similar responses to such stimuli. There is presently no way of knowing if animals develop an ability to think during ontogeny just as humans do, or if their thinking remains at the level of a very young human child.
- In addition, if thinking is the product of learning – a kind of 'cognitive gadget' (Heyes 2018) or software of the brain – then it is a *certain kind* of software or gadget. Otherwise, software such as Tetris would be an example of thinking, and we have reasons to believe that it is not, as such software obediently just steps through its programmed code. Also, in humans most cognitive gadgets, such as literacy and numeracy, do not *in themselves* result in thinking.
- Finally, if thinking is a learnt cognitive gadget, it needs *compatible hardware* – a brain that can incorporate cognitive gadgets. We have reason to believe that this capability of learning not just factual knowledge but to also learn thinking tools is uniquely human. For example, almost all humans, but no other animals – including chimpanzees, bonobos and gorillas – are able to learn language. The currently most promising understanding of what hardware that is missing is that only human brains can decode and utilize general temporal stimulus sequences (Uddén & Bahlmann, 2012; Ghirlanda et al., 2017; Lindenfors 2019; Enquist et al., 2023).

Proposition: Downwards Causation

Higher order thought processes can in themselves have causal effects on the brain, or phrased differently, not all aspects of thinking bubble up involuntarily from our subconscious. To see why this is so, consider the following.

I can change your mind. I can carry out this feat by transmitting information content from my brain to yours, over some physical medium such as sound waves, printed-paper pages, or – most probably – contrasting pixels on an electronic screen. Crucially, however, what changes your mind is not the medium but the *information content* carried by it – its semantic *meaning*. Meaning, in this context, is to be understood as semantic information processing with consequences; and the consequences, in this context, are that this semantic information processing changes your mind.

The meaning is the same whether you listen to or read these words. Although information is encoded physically and fully dependent on physical representation (Landauer, 1996), information content is not itself physical, but exists in abstract form, as thoughts or ideas, 'that what is signified', 'interpretant', or 'memes', if you will. Theories of meaning is a whole field of philosophy that is too vast to summarize

here, however. I will adhere to a simplified definition of meaning: *semantic information processing with consequences*.

There is a difference between the letters in a book and the story, between the zeroes and ones in a picture file and the image, between the notes of an opera and the music itself. All these examples are comparable to the relationship between speech or writing with that which is spoken or being signified. This reasoning follows the semiotic theory of Charles Sanders Peirce, that distinguishes between objects (the actual thing), signs (the representation of that thing), and interpretants (the meaning of the sign) (Peirce, 1998, p. 478). Information content is encoded in the medium, but it isn't the medium – information content is reliant on the material world to be represented but is in itself substrate independent; it transcends a mere explication of the physical properties of the material. Again, semantic information processing can be said to have 'meaning' when it has some consequence in the world.

It follows that – should I succeed in changing your mind – this does not happen through direct manipulation of your brain. I do not invasively add, remove, weaken, or strengthen synapses in your brain's neural network. Even if this is what eventually happens, on a biological/physical level, such changes are consequences of the information content transferred, not its cause.

When you have changed your mind, this does similarly not result in awareness of synapses having been added, removed, strengthened, or weakened. Instead, the awareness is that an idea, understanding, sign, or 'meme' has emerged, been transmitted, refuted, or changed; that your brain's information content has been altered.

These are not controversial statements; in fact, they are trivial. But here is a comparable, more problematic proposition: You can change *your own* mind, through inner manipulation of information content encoded in language. The proposition is that manipulating meaning through manipulating symbols – thinking – causes changes in brain states.

This assertion is more problematic because it implies that internal information processing on a higher level of abstraction (e.g., complex calculations, philosophical deductions, scientific reasoning) causes brain processes on a lower level (neurons firing and synapse-connections changing) instead of causality being the reverse. Or, phrased differently, if higher level information processing changes lower-level brain states, then causality can flow downwards, from higher levels of abstractions (meaning) to lower (brain states).

But is this really what happens when you change your own mind? Isn't it the other way around, that information processing on a lower level generates thoughts, that your higher level though processes are just 'informed' of lower-level thought processes?

One type of causality can at least be ruled out. It cannot be the case that higher level information content changes first and synapse connections later. This would imply that some information processing is carried out independently from the brain, perhaps in a separate 'something' such as a hypothetical soul that can process information independently and then inform the brain after the fact, through for example adding, removing, weakening, or strengthening synapses. I will not be considering this supernatural proposition here.

Instead, I suggest that internal information processing on a higher level of abstraction can cause changes in brain states. This would be indicated by brain states changing simultaneously as the higher-level information content, though the process is driven by the higher-level information processing. I will argue for such downwards causality on logical grounds.

Note, however, that even if this proposition is true, the causal arrow of brain activity is more often expected to point upwards than downwards, from lower-level brain state changes to higher-level thought. For example, during certain brain operations, doctors can stimulate the brain with electrodes in patients and thusly cause conscious experiences. This experience can then be related back to the doctor using language (e.g., Blanke et al., 2002).

The proposition considered here is also not that all, or even most, causality in the brain flows downward. On the contrary, it is well known that the overwhelming majority of the brain's activities are sub-conscious. It has been estimated that the conscious mind has a capacity to process about 50 bits per second (Zimmerman 1987 – the exact number will of course depend on various assumptions and would be expected to vary somewhat depending on individual and task) and that languages universally transmit information at about 39 bits per second (Coupé et al., 2019). For comparison, the senses transmit about 11 billion bits per second to the brain (Zimmerman, 1987). It follows that most information processing handled by the brain therefore must be sub-conscious. The proposition explored here is instead that an *important part* of information processing has downwards causality.

This proposition of downwards causality is empirically testable. Compare the situation to Benjamin Libet's series of experiments where subjects were asked to report on their decision to push a button. Libet recorded information processing occurring *before* the decision, indicating that non-conscious processes gave rise to the experience of having reached the decision to push the button (Libet, 2004). Brain processes first, awareness of the decision later – causality flows upwards.

If information processing on a higher level of abstraction causes simultaneous changes on lower levels this entails that a comparable experiment would be unsuccessful if carried out on decisions arrived at through higher level thought processes. Instead, the decision and the change in brain states would occur simultaneously. Or, stated differently, brain states would be selected based on their information content on a higher level of abstraction/their meaning (idea/meme/signifier). Thinking would change the sub-conscious.

A simple example of when higher level thought processes have downwards causality is the difference between multiplying 4×6 as compared to 342×226. Most of us know from memory that the solution to the first multiplication is 24. However, even though solving the second multiplication is not complex if you have pen and paper, for many of us it involves substantial effort to carry out in our heads; to follow rules laid out during our primary school years, keeping separate products in memory, and adding products over several steps. When we force ourselves to carry out such multiplications, the results do not bubble up involuntarily from our sub-conscious. Instead, we willfully impose the algorithms on abstract information and our uncooperative brains.

Similarly, compare recognizing a common word like ‘apple’ when you read it, to trying to decipher a complex scientific term you’ve never encountered before, such as ‘polytetrafluoroethylene’. You might need to break down this novel word into its component parts, consider its roots, or even sound it out syllable by syllable, to make sense of it.

As another example, consider a syllogism of this type:

- All men are mortal.
- Socrates is a man.
- Therefore, Socrates is mortal.

In this form, the conclusion (‘Therefore, Socrates is mortal’) is a necessary consequence of the two premises above it – it follows logically. Once you have learned to arrive at a conclusion in this manner, you can apply the same rule to any syllogism, or symbolic representation of such a syllogism. When doing this, information manipulation occurs in the brain on a higher abstraction level. It is not enough to manipulate the words separately or consider their individual placements – the same syllogism can be expressed with other words placed in another sequence in another language, or in formula form. The essential information needed to ‘solve’ a syllogism lies in the meaning encoded in the two premises (are they valid and relevant?) and in the logical relationship between the meaning they contain.

Similarly, consider applying theoretical scientific knowledge to draw conclusions:

- All objects in a vacuum fall at the same rate, regardless of mass.
- A feather and a hammer are dropped in a vacuum.
- Therefore, the feather and the hammer will hit the ground at the same time when dropped in a vacuum.

This involves grasping the abstract concept of gravitational acceleration in a vacuum, rather than specific properties of feathers or hammers.

Solving complex multiplications, decoding complex terms, understanding syllogisms, and using scientific knowledge to draw conclusions must occur simultaneously in the mind (internal information manipulation) and in the brain (synapses being weakened or strengthened). But the *cause* of events involves information processing on a higher level of abstraction – the solutions to complex problems cannot bubble up from our sub-conscious as the problems are encoded in information contained on a higher level of abstraction.

From this it is reasonable to infer the conclusion that there exist some cognitive gadgets in our mind (multiplication, decoding of complex terms, logic, and scientific understanding, as exemplified here) where higher level learnt abstract processes cause changes in lower-level brain states. Note that these processes are goal oriented, that we embark on them for specific willfully imposed purposes. However, note also that the suggestion that the brain is ‘programmed’ by information transmitted verbally has been deemed (too) reductionistic (Brette, 2022).

Proposition: Natural Selection & Cultural Selection

To make my following point about thinking, I will reason by analogy with natural selection. As stated above, information content is carried by physical media, but not identical to the media itself. Correspondingly, genes carry information content (see Hoffmeyer & Emmeche, 1991 for an extensive discussion of this proposition). At its most basic, genetic information consists of templates for RNA (and in extension, often for proteins) or modifiers of other genes. On another level, genetic information carries ingredients and blueprints for organism *traits*, to be realized as physical or mental traits of the organism itself (e.g., 'wings', 'alertness'), in relation to others ('monogamous', 'social'), or even outside the organism ('weaver-bird nest', 'beaver dam'). Because the processing of genetic information has consequences in the world, genes carry meaning, as defined above.

DNA (or RNA) *without* meaningful information content reacts with other chemicals in the environment according to the same natural principles that govern all chemical reactions. In such a case, to know why a specific number of copies of a specific DNA-strand exists, it is sufficient to know what chemical substances DNA is made of, the structure of the specific molecule, and its history – i.e., to know what Aristotle termed its material, formal and efficient causes (Falcon, 2023).

DNA (or RNA) *with* meaningful information content similarly reacts with chemicals in the environment according to the same natural principles that govern all chemical reactions. However, in such a case it is not sufficient to know the chemical composition and structure of DNA and the chemical processes that gave rise to it in order to understand why a specific number of copies of a certain strand exists. The information content carried by the DNA has added an extra explanatory level that is not present without this information content. (The fact that many genetic products function to regulate and uphold the genetic machinery does not change this general argument.)

More specifically, meaning seems to equip each gene with a final (Falcon, 2023), ultimate (Tinbergen, 1963) cause (Deacon, 2011): to result in some characteristic of some organism that furthers that organism's survival and/or reproduction, or, more specifically understood, to further the copying of the gene itself. Note, however, that the same observation can be stated without invoking purpose: the meaning that exists has *previously* provided characteristics of organisms that have furthered their survival and/or reproduction.

The frequency of a gene in a population depends to a large degree upon the effects of that gene's products on the organism in its interaction with its environment. Or, stated differently, genetically encoded traits have different effects in different organisms and in different environments, factors completely unrelated to the chemical or physical characteristics of the DNA itself. The number of times a specific gene will be replicated is instead mainly determined by the information content carried by the gene, i.e., the gene product in relation to its environment.

It follows that causation in natural selection flows downward, from the trait encoded by the information carried by each gene to gene-frequency; from natural selection to replication. The higher-level phenomenon of natural selection is not deducible solely from the chemistry and physics of genes.

Note the similarity in this regard between meaning carried by genes and meaning carried by language (e.g., Dawkins, 1976; Dennett, 2017). Noise without semantic information content reacts with the environment according to the same natural principles that govern all reactions to noise in nature. In such a case, it is sufficient to know that noise is compression waves moving through air (i.e., the chemical and physical properties of air) and the processes that gave rise to it (its history), in order to understand all there is to know why a specific noise exists and interacts with the environment the way it does.

Linguistic information content with meaning similarly reacts with the environment according to the same natural principles that govern all reactions to noise. However, in such a case it is not sufficient to know the chemical and physical properties of sound waves travelling through air and the processes that gave rise to it to understand why the noise exists and interacts with the environment the way it does. The meaning carried by language has added an extra explanatory level, a dimension not present without meaningful information content.

As in the case with genes, meaning seems to equip utterances with a final, ultimate cause (Deacon, 2011): in this case to result in some changed opinion or behavior in a recipient organism that furthers the subsistence and/or transmission of that information. This potential final cause can be realized more or less successfully, resulting in more or less numerous copies of the specific information content in question. The ultimate cause – the purpose of any piece of linguistic information – is not present without information content carried by language. Again, however, the same observation can be stated without invoking purpose: the linguistic information content that previously has provided subsistence and/or transmission of that information. (Note here too that many language products function just to regulate and uphold the language machinery in itself. Again, this does not change the general argument.)

An empirically deduced characteristic of all living things is that they contain self-replicating, information carrying molecules: DNA or RNA. In principle, however, as for all information, the choice of information medium is arbitrary. To see that this is so, imagine a set of micro-robots manufacturing proteins and RNA from genetic code fed to them from a computer; the computer in turn prompted by chemical signals from the cell. All functions can (in theory) be carried out by such micro-robots as by DNA. The crucial ingredient from the point of view of the cell is the information content of the DNA, not the DNA itself, and that the information carried has the same consequences.

For semantic meaning, the case is similar. Again, the choice of information medium is arbitrary. To see that this is so, imagine John Searle's (1980) 'Chinese Room' thought experiment – where a person is locked in a room replying to written Chinese symbols according to instructions written in an instruction book – but this time envision a *population* of Chinese rooms exchanging information.

Here, the processes carried out by each Chinese Room is comparable to those of Chinese persons having written conversations. The exact same functions are fulfilled by the Chinese rooms as by Chinese persons, and the conversation can thus be continued, even though there is no single part of any 'Chinese Room' that knows Chinese. The retention and / or transmission frequency of meaning in a 'Chinese Room' population is in this scenario only dependent on the meaning, not on the characteristics of

the Chinese rooms themselves. The ultimate, final cause for how each Chinese room interacts with information content is encoded in the information content carried by the Chinese language.

Note here that all Chinese-speaking persons functionally are versions of 'Chinese rooms' themselves, as they all consist of parts that do not understand Chinese. Also, the information content transferred can be of the kind that rewrites the instruction book itself, thus learning new methods for how to handle information. This is, however, only possible in cases where something similar to an 'instruction book' exists, i.e., in systems utilizing some form of a symbolic system for information processing. Thus, this argument is particular to language.

If the argument presented here has any validity, the driving force of life emerged with information-carrying molecules. By analogy, the driving force of thinking emerged with language. Retention and/or transmission of information exhibit downwards causality in both systems.

To generalize, the selection of genes on basis of the interaction of their information content (what they code for) with the environment and the selection of brain states on basis of the interaction of their information content (what they *mean*) with the environment, are two comparable processes.

Hypothesis: from Language Processing to Thinking

How then to get from a bare brain to thinking during childhood development? What follows are some hypotheses on this point. As a reminder, the proposal under consideration is that thinking itself is a product of verbal programming – a thought-program running in the brain – akin to a 'virtual machine' as suggested by Dennett (1991), or a 'cognitive gadget' proposed by Heyes (2018). Thinking is, in this regard, internal symbolic responses to stimuli that are either intrinsic or extrinsic, and computational procedures that operate on these internal symbolic representations. Note also that the blueprint of the gadget itself is left unspecified – the argument here is about its existence.

The kind of symbol manipulation that language represents could theoretically emerge in any medium, but currently exists – as far as we know – only in human brains. A dog cannot change your mind and you cannot change a dog's mind – you can only make it act differently through classical or instrumental conditioning (Heyes 2012; Enquist et al., 2023). There is now novel empirical evidence that this inability of animals to learn language has to do with limitations to decode and utilize temporal stimulus sequences (Ghirlanda et al., 2017; Lindenfors 2019; Enquist et al., 2023).

There exist other types learning besides symbolic, such as classical and operant conditioning, and transferal of skills from one cognitive domain to another. Only language, however, can convey truth propositions, what Peirce terms 'dicisigns'; statements that can be either true or false (Stjernfelt, 2014), which is why it is of particular interest in the case of computational procedures that operate on internal symbolic representations.

Whatever the cause of animals' inability to learn language, there exists a qualitative difference between 'ordinary' sensory processing, as occurs in all neurons and

neuronal bundles in the animal kingdom, and language processing, as occurs only in humans. In the first case, the information processed is provided by nature through sights, sounds, smells. In the second case, language processing enables the processing of abstract information through the manipulation of symbols (signifiers, memes) and concepts. Language is one abstraction level up from ordinary brain processes.

This distinction between two types of information processing is important, because symbolic manipulation means that language can function as a programming language, making possible not only the transmission of information per se, but the transmission of information processing *algorithms*, making programming the mind possible. This is not meant as a metaphor, but as a factual statement – language and programming languages are functionally equivalent in this regard. Human upbringing is, according to this description, a long programming process where factual information, thought tools and algorithms are inscribed into the brain over many years (but see Brette, 2022). To bring it back to the Chinese room population-thought experiment, language brings with it the possibility to write and edit the instruction books in the Chinese rooms.

Similar viewpoints to those stated here have been extended and discussed by programmers that work on artificial intelligence. For example, Sloman and Chrisley (2003) have proposed something they call "Virtual machine functionalism", where mental states and processes are to be understood as operations within a kind of "virtual machine" implemented on the brain's neural hardware. Virtual machine functionalism simply suggests that the mind operates like a virtual machine running on the brain's neural hardware. Notably also, current Large Language Models have gained capabilities very similar to reasoning from learning language processing.

More recently, philosopher Dennett (2017), has argued that culture and evolution are intertwined, with memes (units of cultural evolution) playing a pivotal role in shaping human minds. Doing this, he challenged traditional views of consciousness, suggesting that it's not a single thing but a complex assembly of numerous brain processes. Central in Dennett's thinking is the idea of "competence without comprehension" – abilities to perform specific tasks without necessarily having comprehension or conscious awareness of those tasks.

In the context of evolution, Dennett suggests that natural selection can produce organisms that are highly adept at survival and reproduction without these organisms understanding or being conscious of the strategies they employ. For example, a spider can weave a complex web without understanding the principles of web design; ants construct ant nests without comprehending nest design. In AI, this idea is exemplified by machine learning algorithms that can perform complex tasks, like image recognition, generative language models, or playing chess at a high level, without any comprehension (that we know of) of the task they are performing.

We humans demonstrate competence without comprehension in many everyday activities. For instance, many of us can speak our native language fluently without understanding or being able to explain all grammatical rules governing that language. Dennett argues that our capacity for symbolic reasoning and our ability to use and understand complex language structures have significantly influenced the evolution of our minds. He suggests that human brains have become adept at hosting and propagating memes – ideas, behaviors, or styles that spread within a culture. Meme-gene

co-evolution has, according to Dennett, played a significant role in the development of human consciousness.

Dennett's concept is significant as it suggests that comprehension, consciousness, or understanding are not prerequisites for high-level competence. This has implications for how we think about both natural intelligence in humans and other animals, and artificial intelligence in machines. It also raises philosophical questions about the nature of consciousness and its role in cognitive processes. This concept challenges traditional views that associate high-level competence with a corresponding level of conscious understanding or reasoning.

Dennett's work, including his views on competence without comprehension, is tied to his eliminativist approach to consciousness. He suggests that phenomena like qualia, the subjective aspects of consciousness, are illusory intentional objects of our introspective beliefs. This aligns with his view that complex functions like consciousness can emerge from systems that are competent in certain tasks without necessarily having an understanding of those tasks.

Heyes (2018) has argued that language provides a possibility to improve, change, or develop your own thinking. Contrary to the prevailing view in evolutionary psychology that sees the human mind as a collection of cognitive instincts shaped by genetic evolution over long time periods, Heyes introduced the idea of "cognitive gadgets." She argued that humans possess special-purpose organs of thought that are constructed during development through social interaction. However, these cognitive gadgets are products of cultural evolution, not genetic evolution. This means they can develop and change more rapidly and flexibly than cognitive instincts. Heyes' theory emphasizes the significant role of cultural transmission and learning in shaping human cognitive abilities.

It has also been pointed out by AI-theoretician Hofstadter (1979) that symbol systems that are applied on themselves can illuminate the system's own qualities in a manner that is not possible without such a symbol system. Hofstadter's central concept revolves around the idea of "strange loops:" self-referential structures or patterns that loop back on themselves within hierarchical systems. As one progresses through the levels of a system, such a "strange loop" can bring one back to the starting point, creating a paradoxical situation. Hofstadter suggests that consciousness, meaning, and identity might emerge from similar recursive processes, where elements refer back to themselves. Symbolic self-reference can, according to Hofstadter, in this way be the explanation of self-awareness, the ability to think about one's own thinking.

In a similar manner, the psychologist Julian Jaynes has controversially suggested that self-awareness emerged when people became able to start thinking about their own thoughts – meta-consciousness (thinking about thinking; consciousness about your own consciousness). In his book *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (1976) he introduced the idea of the "bicameral mind," a non-conscious mentality prevalent in early humans that relied on auditory hallucinations. He suggested that consciousness is a learned behavior rooted in language and culture, rather than being innate. Jaynes hypothesized that the transition from the bicameral mind to consciousness occurred around the 2nd millennium BCE, triggered by the breakdown of the bicameral system.

Terrence Deacon's books *The Symbolic Species* (1997) and *Incomplete Nature* (2011) explores similar themes. Merging insights from neurobiology, evolutionary theory, linguistics, and semiotics, Deacon drew the conclusion that the unique human capability for symbolic thought and language co-evolved with the brain. A central theme of the first book is the interdependence between symbolic thought and language, presenting a chicken-and-egg dilemma: while language is the medium for symbolic thought, mastering language would seemingly necessitate prior symbolic thinking capabilities. Deacon resolves this puzzle by suggesting that language and symbolic thought evolved in tandem. In his second book, Deacon further argues that just as the notion of zero revolutionized mathematics, considering life and mind in terms of constraints (what is absent) can help address the mind-body problem.

Language as information carrier is a fundamental entity. To speculate on the evolution of language is a topic outside the scope of the current article, but for the current purpose it helps to break apart some ways in which language conveys information. Language transfers information on three levels: (1) *that* something is said (indicating that both speaker and listener possess the biological ability to use language), (2) *how* it is said (what language is used, indicating a common culturally evolved symbol system), and (3) *what* is said (the actual information content of each utterance – its meaning). Even if 'what is said' may be manipulative and carry false information, one is always bound to cooperate honestly on 'how it is said' and 'that it is said' – to not cooperate on how and what is said results in incomprehensible statements, not deception (Lindenfors, 2013).

Though there are anecdotes of intentional deception in primates (e.g., de Waal 1992), these are just that – anecdotes. It is theoretically almost impossible to envision the evolution of a trait that immediately can be misused for deception within species, and some evidence instead suggests that deception may be a derived function of language (Oesch 2016).

However language evolved, if programming of the brain during childhood is what gives rise to self-awareness, then self-awareness is an epiphenomenon of linguistic information manipulation. Such a 'programmed self' is, as we have seen, falsifiable, has a set of interesting implications, and solves some conceptual problems in the science of thinking.

According to the theory is outlined here, a human 'I' would consist of three processes, the last of which gives rise to thinking.

- **Biological information processing**. Naturally selected factual and procedural knowledge that we are born with, hard encoded in our brains. This kind of information processing occurs in all living creatures with neurons. This is the type of process that gives rise to the sensation of pain.
- **Learnt information processing**. Factual and procedural knowledge selected through interactions with the environment, learnt through associative learning and chaining mechanisms. This kind of information processing occurs in all animals that can learn through classical and operant conditioning. This is the kind of process that gives rise to pain avoidance.
- **Thought, or internal symbolic information processing**. Factual and procedural knowledge selected through reflection, learnt via language. This kind of informa-

tion processing occurs only in animals that can learn language – humans. Since it is learnt via language, it is a culturally evolved trait. This is the type of process that gives rise to an understanding of pain avoidance.

Note that all three processes affect the same organ – the brain. This means that which of the brain's products that comes from what process may be extremely difficult to sort out. Note also that genetic inheritance and basic associative learning mechanisms exist in the simplest single-celled animals all the way up to the animals we consider most intelligent, such as chimpanzees, bonobos, dolphins, and corvids. It is sometimes claimed that there exists one more level of information processing in animals: non-linguistic contemplation of problems in, for example, apes and corvids, but this is now strongly disputed (Lind et al. 2009; Ghirlanda et al. 2017, Lindenfors 2019; Enquist et al., 2023; Lind et al., 2023).

Only one species seems to be able to learn to incorporate and utilize linguistic information in a manner that fundamentally alters our behavior and capacity for social learning: humans. Once brains can incorporate software, a new selection process begins on the software itself: cultural evolution, where the survival and reproduction of 'memes' matter more than the survival and reproduction of their carriers.

Potential Consequences of this View of Thinking

If the minds 'I' is verbally encoded software, reacting to input from and outputting commands to the hardware, then there really *is* a Cartesian theatre, with a software 'I' as the onlooker – an observer gadget (Heyes, 2018) or virtual machine Dennett (1991). The homunculus in the machine is, in this view, a 'self-awareness program' rather than just another hardware entity of the same kind as the one that was to be explained in the first case. In this view, there thus exists a real and tenable difference between *res cogitans* (the realm of thought) and *res extensa* (the realm of extension), similar to Cartesian dualism, in an analogous manner as there is a difference between information content/meaning and the information carrying media.

Thinking, then, would be the result of the processing of a specific (but yet unspecified) brain software, not the software itself, but its non-material consequence. (Software should here be understand in a functional sense – information incorporated and processed in the brain – not in the literal sense with a one-to-one correspondence to a von Neumann architecture computer with transistors, logic gates, etc.)

Further, if the minds 'I' is the result of such linguistically infused software, this would explain the sense of unity that self-awareness provides that comes from the 'I' actually being a coherent unit and not 'only' a collection of sensory impressions and reactions on these. To a large degree, the brain runs in parallel, and its functions are localized in a number of different places. Thought processes are not physically centralized – there is no executive center. Such a collection of activities could, however, be collected through a thought program where the program is the 'onlooker' – a thought program able to handle about 50 bits per second.

Self-awareness is 'knowing that one knows'. It is the difference between experiencing the redness of red, the taste of great beer, and just neuronal signals encoding

'red' and 'tasty'. If there is no encoded self-awareness program in a brain, then there is no software receiving the signals. Since animals cannot master language, they cannot have their brains programmed and thus cannot possess any thought processes of this kind. If this software view of thinking is correct, there may therefore be nothing it 'is like to be a bat' – or at least not something the bat would be aware of as how it is to be. (I. e. *nothing or no one knows* what it is like to be a bat.) In fact, in extension it would also mean that nothing or no one knows what it's like to be a chimpanzee or a small child. (This may be an emotional reason to resist the proposition under consideration.)

What follows is also that if thinking is an ability that is encoded verbally, learnt via language; not innate, but gained throughout childhood. This would explain why we have no memories from when we were very small children – we did not possess the 'gadget' that could contextualize experiences and encode them in a meaningful and useable way into the brain. The abilities we retain from early childhood are knowledge and competences that we either are born with (such as breathing) or lower level information processing capabilities that learnt through associative learning (such as walking or cycling), not memories dependent on ideas or memes dependent on higher-level information processing.

If the 'minds I' is software, programmed during childhood, then there is an actual gap between humans and other animals. Why animals cannot handle language and thus cannot download 'thought-gadgets' into their brains may spring from a small difference of competence in temporal sequence handling (Ghirlanda et al., 2017; Lindenfors 2019; Enquist et al., 2023) and/or other differences between humans and animals. The gap, however, would not emerge until humans have had their brains programmed, after mastering enough language.

In the software view, brain parts can have their own self-awareness if they can run the self-awareness software on their own. Brain parts can also run this program together, in which case the program as a whole is the 'self'. This could be the explanation of the oddities when observing split-brain patients, who appear to have two minds running in parallel – the program may run well on just half a brain (but not exactly equally well in parts less involved in language processing). Note also that a thought program, as proposed here, can run on any compatible hardware – on any suitable medium. There is no need to run the program on meat.

Further, if the software view of thinking is true, this may be a possible explanation of the odd disorder of 'blindsight', where the visual system is intact but the patient nevertheless experiences a loss of vision. What is defect here may be either the connection between the software and the hardware, or a glitch in the software itself. This malfunction may be due to errors either in the software or in the hardware. From a software view of self-awareness, we would expect uniquely human psychological problems stemming from software errors, in addition to psychological problems stemming from hardware errors that should also occur in animals.

Author Contributions PL carried out all work associated with this manuscript.

Funding Open access funding provided by Stockholm University.
No funding was received for conducting this study.

Declarations

Competing interests The authors declare no competing interests.

References

Blackmore, S. (2003). Consciousness in Meme machines. *Journal of Consciousness Studies*, *10*, 19–30.
Blackmore, S., & Troscianko, E. T. (2018). *Consciousness: An introduction*. Routledge.
Blanke, O., Ortigue, S., Landis, T., & & Seeck, M. (2002). Stimulating own-body perceptions. *Nature*, *419*, 269–270.
Block, N. (1995). The mind as the Software of the brain. In E. Smith, & B. Osherson (Eds.), *Invitation to Cognitive Science*, *3*, Thinking, 377–425.
Brette, R. (2022). Brains as computers: Metaphor, analogy, theory or fact? *Frontiers in Ecology and Evolution*, *10*, 878729.
Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 200–219.
Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, *5*, eaaw2594.
Dawkins, R. (1976). *The selfish gene*. Oxford University Press.
Deacon, T.W. (1997). The symbolic species: *The co-evolution of language and the brain*. WW Norton & Company.
Deacon, T. (2011). *Incomplete nature: How mind emerged from Matter*. W.W. Norton & Company.
Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Co.
Dennett, D. C. (2017). *From Bacteria to Bach and back: The evolution of minds*. WW Norton & Company.
De Waal, F.B. (1992). Intentional deception in primates. *Evolutionary Anthropology: Issues, News, and Reviews* 1:86-92.
Enquist, M., Ghirlanda, S., & Lind, S. (2023). *The human evolutionary transition: From Animal Intelligence to Culture*. Princeton University Press.
Falcon, A. (2023). Aristotle on Causality. *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), EN Zalta & U Nodelman (Eds.) https://plato.stanford.edu/archives/spr2023/entries/aristotle-causality
Fodor, J. (1975). *The Language of Thought*. MIT Press.
Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: A divide between humans and other animals? *Royal Society Open Science*, *4*, 161011.
Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.
Hoffmeyer, J., & Emmeche, C. (1991). Code-Duality and the Semiotics of Nature. In M. Anderson, & F. Merrell (Eds.), *On semiotic modeling* (pp. 117–166). Mouton de Gruyter.
Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Basic Books.
Jaynes, J. (1976). *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston: Houghton Miffli
Landauer, R. (1996). The physical nature of information. *Physics Letters A*, *217*, 188–193.
Libet, B. (2004). *Mind time: The temporal factor in consciousness, perspectives in cognitive neuroscience*. Harvard University Press.
Lind, J., Ghirlanda, S. & Enquist, M. (2009). Insight learning or shaping? *Proceedings of the National Academy of Sciences* 106: E76-E76.

Lind, J., Vinken, V., Jonsson, M., Ghirlanda, S., & Enquist, M. (2023). A test of memory for stimulus sequences in great apes. *Plos One*, *18*(9), e0290546.
Lindenfors, P. (2013). The green beards of Language. *Ecology and Evolution*, *3*, 1104–1112.
Lindenfors, P. 2019 Det kulturella djuret: Om människans evolution och tänkandets utveckling. Ordfront, Stockholm.
Oesch, N. (2016). Deception as a derived function of language. *Frontiers in Psychology* 7, p.220523.
Peirce, C. S. (1998). *The essential Peirce. Volume 2. Eds. Peirce Edition Project*. Indiana University Press.
Putnam, H. (1980). Brains and behavior. *Readings in Philosophy of Psychology*, *1*, 24–36.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417–424.
Sloman, A., & RL Chrisley (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, *10*, 133–172.
Stjernfelt, F. (2014). *Natural propositions: The actuality of Peirce's doctrine of Dicisigns*. Docent.
Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433.
Uddén, J., & Bahlmann, J. (2012). A rostro-caudal gradient of structured sequence processing in the left inferior frontal gyrus. *Philosophical Transactions of the Royal Society Series B*, *367*, 2023–2032.
van Gulick, R. Consciousness. (2022). The Stanford Encyclopedia of Philosophy (Winter 2022 Edition). Edward N. Zalta & Uri Nodelman (eds.). https://plato.stanford.edu/archives/win2022/entries/consciousness
Zimmerman, M. (1987). The nervous system in the context of information theory. In R. F. Schmidt, & G. Thews (Eds.), Human physiology (pp. 166–173). Springer