Philosophy and Technology manuscript No.
(will be inserted by the editor)

Zombie mouse in a Chinese room

Slawomir J. Nasuto · J. Mark Bishop · Etienne B. Roesch · Matthew C. Spencer

the date of receipt and acceptance should be inserted later

Abstract John Searle's Chinese Room Argument (CRA) purports to demonstrate that syntax is not sufficient for semantics, and hence that because computation cannot yield understanding, the computational theory of mind, which equates the mind to an information processing system based on formal computations, fails. In this paper, we use the CRA, and the debate that emerged from it, to develop a philosophical critique of recent advances in robotics and neuroscience. We describe results from a body of work that contributes to blurring the divide between biological and artificial systems: so-called animats, autonomous robots that are controlled by biological neural tissue, and what may be described as remote-controlled rodents, living animals endowed with augmented abilities provided by external controllers. We argue that, even though at first sight these chimeric systems may seem to escape the CRA, on closer analysis they do not. We conclude by discussing the role of the body-brain dynamics in the processes that give rise to genuine understanding of the world, in line with recent proposals from enactive cognitive science[1].

Slawomir J. Nasuto
Univ. Reading, Reading, UK, E-mail: s.j.nasuto@reading.ac.uk

J. Mark Bishop
Goldsmiths, Univ. London, UK, E-mail: m.bishop@gold.ac.uk

Etienne B. Roesch
Univ. Reading, Reading, UK, E-mail: contact@etienneroes.ch

Matthew C. Spencer
Univ. Reading, Reading, UK, E-mail: matthew.spencer@pgr.reading.ac.uk

[1] In this work the term enactive cognitive science will be used to delineate theoretical approaches to cognition that emphasise perception as action encompassing, for example, Gibson's "ecological approach"; Varela, Thompson and Rosch's "embodied mind"; Nöe's "action in perception" and O'Regan and Noë's "sensorimotor account of vision".

1 Searle's Chinese Room Argument

1.1 The CRA in a nutshell

In developing the Chinese Room Argument in his paper 'Minds, Brains and Programs' (MBP) [Searle(1980)] John Searle gave birth to one of the most notorious debates in the history of philosophy of mind [Cole(2009), Preston and Bishop(2002)]. Searle's argument refutes the strong claim of artificial intelligence, which he coined "Strong AI", that of creating a truly intelligent computational device, demonstrating machine understanding [Searle(1980), p. 417]. This lasting debate has important consequences for cognitive science in general, and the computational theory of mind in particular, which equates the mind to an information processing system based on formal computations [Fodor(1975), Fodor(1987), Putnam(1988)]. If Searle's position in the debate is correct, it also shows the inadequacy of any purely behavioural procedure (e.g. Turing's 'test') to identify intelligence[2].

In 1977, Schank and Abelson published information on a script[3] based software system that they had developed to answer questions on simple stories [Schank and Abelson(1977)]. The program takes as input a simple story and, using sets of rules, heuristics and scripts, is able to infer answers to questions about the story posed by an operator. Searle gives the example of a story depicting a man entering a restaurant, ordering a hamburger and storming outside the restaurant disappointed. If asked "Did the man eat the hamburger?", the program would unequivocally answer "No, he did not", based on its script outlining the typical expected behaviour exhibited by customers visiting restaurants.

In the CRA, by demonstrating that syntax is not sufficient for semantics, Searle responds to the claim that the appropriately programmed computer genuinely 'understands' the data it processes; for example, the claim that Schank and Abelson's software actually understands the text it processes. To this end, Searle described a thought experiment in which he was locked in a room, and provided with 'a large batch of Chinese writing'. As Searle does not speak a word of Chinese, to him these Chinese ideographs were "just so many meaningless squiggles" (ibid. p. 418). Searle is then presented with a second batch of Chinese script together with a set of rules [which Searle can understand] for correlating the second batch with the first batch. Some time later Searle is subsequently given:

> ... a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch

[2] In what has become known as the standard interpretation of the Turing test, a human interrogator, interacting with two respondents via text alone, has to determine which of the responses is being generated by a suitably programmed computer and which is being generated by a human; if the interrogator cannot reliably do this then the computer is deemed to have passed the Turing test.

[3] In their work Schank and Abelson used *scripts* to specify a detailed description of stereotypical events unfolding through time in given contexts.

> with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch.

Unbeknownst to Searle:

> ... the people who are giving me all of these symbols call the first batch "a script", they call the second batch a "story" and they call the third batch "questions". Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions" and the set of rules in English that they gave me, they call "the program".
>
> Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view, that is, from the point of view of somebody outside the room in which I am locked, my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese.
>
> Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view – from the point of view of someone reading my "answers" – the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.

Thus Searle's rulebook describes a procedure which, if carried out accurately, allows him to participate in an exchange of uninterpreted symbols - squiggles and squoggles - which, to an outside observer, look as though Searle is accurately responding in Chinese to questions in Chinese about stories in Chinese; in other words, it appears as if Searle, in following the rule book, actually understands Chinese, even though Searle trenchantly continues to insist that he does not understand a word of the language.

According to Bishop [Bishop(2004)], "the central claim of the CRA is that computations alone cannot in principle give rise to understanding, and that therefore computational theories of mind cannot fully explain human cognition. [...] And yet it is clear that Searle believes that there is no barrier in principle to the notion that a machine can think and understand; [...] Searle explicitly states, in answer to the question 'Can a machine think?', that 'the answer is, obviously, yes. We are precisely such machines' ".

Searle's "intuition pump", a term coined by Dennett [Dennett(1991)], provoked an intense reaction in the AI community who attempted, but arguably failed, to demonstrate that the CRA was wrong. What have emerged as perhaps the most important criticisms, which Searle anticipated in the original exposition of the CRA, are those Searle termed the "systems reply", the "robot reply" and "the Brain Simulator reply". Even in the original exposition of the CRA, Searle took these criticisms seriously, presciently anticipating several key turns in recent cognitive robotics, AI and cognitive science.

In the more than thirty years since its first publication there has continued to be lively interest in the CRA[4]; as Bishop [Bishop(2009)] observes:

> .. in a 2002 volume of analysis [Preston and Bishop(2002)] comment ranged from Selmer Bringsjord who observed the CRA to be "arguably the 20th century's greatest philosophical polarizer" [Bringsjord(2002)], to Rey who claims that in his definition of Strong AI Searle "burdens the [Computational Representational Theory of Thought (Strong AI)] project with extraneous claims which any serious defender of it should reject" [Rey(2002)]. Nevertheless, although opinion on the argument remains divided, most commentators now agree that the CRA helped shift research in Artificial Intelligence away from classical computationalism (which, pace Newell and Simon [Newell(1976)], viewed intelligence fundamentally in terms of symbol manipulation) first to a *sub-symbolic neural-connectionism* and more recently, moving even further away from symbols and representations, towards *embodied* and *enactive* approaches to cognition. Clearly, whatever the verdict on the soundness of Searle's Chinese room argument, the subsequent historical response offers eloquent testament to his conclusion that '*programs* are not minds'.

This paper is not an attempt to buttress the CRA. In what follows, we will briefly review the 'System' and 'Robot' replies, before introducing a number of successes in a new branch of robotics that contributes to blurring the divide between biological and artificial systems. We aim to use these examples to articulate a response to current trends in cognitive robotics in line with Searle's position as espoused in the CRA. To cut to the gist, we show that *if the CRA holds then it also holds against both cognitive robotics and bio-machine hybrids (animats) [as currently engineered].*

1.2 The systems reply

The systems reply originated from researchers who took a bird's eye view of Searle's thought experiment. To them, understanding does not lie within Searle, but within the system as a whole. That is, the room plus Searle, plus

[4] Cf. [Rapaport(2006)], [Waskan(2005)], [Sprevak(2005)], [J.(2004)], [Freeman(2003)], [Freeman(2004)], [Overill(2004)], [Garvey(2003)] etc.

the rulebook, plus the sheets of paper [on which are inscribed the various squiggles and squoggles], as it is all these things - the system - as a whole that exhibits the responses perceived as Chinese. Searle responds to this by pointing out that, even if he *internalised* [memorised] the rulebook and all the paper squiggles and squoggles, and hence interacted with the native Chinese speakers directly, laboriously following the instructions of the rulebook as memorised, he would still not understand a word of Chinese.

1.3 The robot reply

The robot reply acknowledges that understanding requires some degree of interaction with the world. Proponents of this position extend the CRA to a robot that interacts with the world through actuators, and 'perceives its world' through appropriate sensors. Using a procedure similar to Searle's rulebook, a computer decides the appropriate symbolic control signals (squoggles) in response to the symbolic descriptions of the world (squiggles) that its sensors present; surely in this case the robot, interacting with the world appropriately at all times, could be said to genuinely 'understand'...

However, by making what is now a well rehearsed move, Searle once again claims that as both the squiggles and squoggles remain merely uninterpreted symbols, a series of binary digits, they would remain meaningless to him as he followed the instructions in the rule book and controlled the robot's behaviour; in other words, that Searle in controlling the robot via the rule book would understand nothing of the robot's interactions with the world and thus - as there is nothing a computer would have that Searle, scraps of paper and rule book do not - neither would the robot.

In other words, the situation Searle describes highlights that, in merely executing its control program, it would always fail to obtain the intimate connection with the world required to give rise to genuine intentional states, genuine understanding and genuine meaning.

1.4 The Brain Simulator reply

The third reply assumes an accurate computer model of the neural mechanisms at play in the brain of a native Chinese speaker, as they respond to questions in Chinese about a story in Chinese. Advocates of this move in the debate assert that to deny genuine understanding to a hi-fidelity neural simulation would be tantamount to denying understanding to the native Chinese speakers themselves.

Searle responds to this proposal by suggesting a replacement of the neurons and synapses with a complex functional analogue [of the neural simulation of the native Chinese speaker's brain] constructed from an interconnection of water-pipes and valves, each of which he would activate according to a rulebook upon receiving a specific series of squiggles as input and then, contingent upon

specific flows at the output of this water-pipe network, the rule book would specify which Chinese ideographs to output. Perhaps not surprisingly, Searle once again concludes that, to him [and the network of pipes], the Chinese ideographs remain meaningless.

2 Hybrid systems and levels of embodiments

The replies to Searle's thought experiment describe situations that are both relevant and conceivable: each situation emphasises particular perspectives of the CRA, and could give rise to further investigation in the form of actual physical/biological experiments with tangible implementations. Today's proponents of so-called embodied AI, a field now known as cognitive robotics, take at least some of Searle's comments seriously in that cognitive robotics acknowledges that mere syntactical manipulation of uninterpreted symbols is insufficient for understanding and emphasises the importance of embodiment in intelligent action [Pfeifer(2001)]. This work has employed extremely varied strategies, the success of which, however, remains debatable [Roesch(in press)].

Interestingly, a fourth line of reply to the CRA (Searle calls the 'combination reply') involves a mix of the previous three replies. In this particular situation, one assumes the examination of a robot in the world, operated by a synthetic brain modelled after a native Chinese speaker. Searle agrees with the contenders of this line of thought: "I entirely agree that in such a case we would find it rational and indeed irresistible to accept the hypothesis that the robot had intentionality, *as long as we knew nothing more about it.*." [Searle(1980), our emphasis]. Once we understood that its behaviour was in fact the result of a [very complex] rule book, we would retract our initial hypothesis and deny the system had genuine understanding.

3 Complex rule books

Historically, Artificial Intelligence (AI) practitioners have been incredulous at the extreme simplicity of the low-level rules described by Searle (and actually deployed in a famous experimental 'recreation' of the CRA by Harré and Wang [Harré and Wang(1999)]) that simply 'correlate one set of formal symbols with another set of formal symbols 'merely by their shape', such that typically very trivial combinations of un-interpreted symbols - squiggles - map simply onto others - squoggles. It has always seemed likely to such AI experts that any machine understanding program with a claim to real-world generality would require a very large and complex rule-base (program), typically applying very high-level rules (functions)[5] and make extensive use of internal variables

[5] In contrast to the thirteen basic ideographs deployed by Harré and Wang, IBM's WATSON system - which recently won world wide acclaim as rivalling the greatest human players of the USA TV game show 'Jeopardy' - effectively deployed a complex high-level rule book (literally thousands of complex algorithms working in parallel) on the full gamut of natural human language.

('pieces of paper' onto which the man in the room [Searle] can scribble symbols and define internal 'representations').

However, it is equally clear from 'Minds, Brains and Programs' [Searle(1980)] that Searle intended the CRA to be fully general - applicable to any conceivable [now or future] AI program (grammar-based; rule-based; neural network; Bayesian etc): '*I can have any formal program you like, but I still understand nothing*'. So if the CRA succeeds, it must succeed against even the most complex 'high-level' systems.

So, in a spirit of cooperation (between computer scientists, AI practitioners and Searle) let us consider a more complex formal program/rule book-system which has (as one high-level-rule) a call to, say, Google-translate. We suggest that these 'internal representations' scribbled on bits of paper, used by the man in the room (monoglot Searle), could now maintain [at least partial] interpretations of the [unknown] Chinese text, as 'symbol-strings-in-English'.

In this way it is plausible that, via a process analogous to one's gradual understanding of a Chinese text via the repeated use of a Chinese-English dictionary, the application of [grounded] high-level-rules (Google-translate) to Chinese text would, over time, foster the emergence of genuine semantics and understanding in even a monoglot English speaker like Searle. Because both the rule book and any internal representations the rule book requires (Searle's 'scribbles on paper') are encoded in English, and *ex hypothesisi* Searle brings to the room an understanding of English, we suggest, after Boden [Boden(1988)], that over time this *extended English Reply* would lead to the emergence of genuine semantics for Searle.

But does a computer Central Processing Unit[6] (CPU) really 'understand' its program and its variables [encoded as raw binary data] in a manner analogous to Searle's understanding of his rule book and internal-representations encoded in English? In her 1988 paper (ibid) Maggie Boden suggests that, unlike say the human-driven manipulations of formal logic, it does; because, unlike the rules of logic, the execution of a computer program 'actually causes events to happen (e.g. it reads and writes data [or instructions] to memory and peripherals)' and such 'causal semantics' enable Boden to suggest that it is a mistake to regard [executing] computer programs as pure syntax and no semantics; such a CPU processing Chinese symbols really does have a 'toe-hold' on [Chinese] semantics. The analogy here is to Searle's understanding of the English language rule book and hence the [extended, high-level] English reply holds.

In the 'Philosophical Investigations' Wittgenstein points out [Wittgenstein(1958)] that there is a fundamental difference between normative rule-following and acting in accordance with rules (obeying physical laws). In contrast to Boden, we assert that the execution of a computer program is merely acting in accordance with physical laws; the CPU does not understand its internal-

[6] A CPU is the core component of a computer system that executes program instructions (its algorithm or rule book) by physically, and in most modern computers typically electronically, fetching or storing (reading or writing) them to and from memory and evaluating their coded commands.

representations [as it executes its program and input] any more than water in a stream 'understands' its flow down hill; both are processes strictly entailed by their current state and that of the environment (their 'input').

Furthermore, as Cassirer [Cassirer(1944)] suggested, we do not consider the computer as it executes its program with particular input(s) an 'information processor' with a concomitant 'toe-hold in semantics', because we consider that the [physical] computer does not process rich semantic 'symbols' (which belong to the human realm of discourse), rather mere un-interpreted 'signals' (binary digits [$\pm$5V]): objects devoid of meaning which belong to the world of physics.

All syntax and no semantics[7] - as there is no genuine sense in which the CPU understands its rule book in a manner analogous to Searle's understanding of English, we suggest that a CPU executing its program is simply not analogous to monoglot-Searle's gradual understanding of a Chinese text via repeated use of an English/Chinese dictionary.

To reflect that the CPU merely mechanically transforms the signals it processes we simply insist that the rule book is defined only by syntactical operations (albeit perhaps more complex than the simple 'correlations' originally suggested by Searle and physically deployed by Harré and Wang) and the internal-representations ('scribbles on paper'), must remain doggedly defined by *meaningless* signals (i.e. un-interpreted symbols; aka Searle's 'squiggles and squoggles').

It is clear that, even allowing the rule book to deploy high-level calls to, say, Google-translate, no understanding of the underlying Chinese text can ever emerge because the 'internal-representations' Searle is forced to manipulate remain uninterpreted ('squiggles and squoggles'). The process is analogous to monoglot Searle's frustrated attempts to understand an unknown Chinese text using only a Chinese dictionary.

4 On epistemology and ontology

In further reflection on the System reply to the CRA, John Haugeland [Haugland(2002)] asks why we should accept Searle's conclusion that the internalised Chinese subsystem doesn't understand Chinese given that its responses to the Chinese questions are correct and indistinguishable from those a native Chinese speaker may give:

> "What we are to imagine in the internalization fantasy is something like a patient with multiple personality disorder. One "personality", Searle, is fluent in English (both written and spoken), doesn't know a word of Chinese, and is otherwise perfectly normal (except that he has the calculative powers of a mega idiot savant). The other ostensible personality - let's call him Hao - is fluent in Chinese (though only writ-

[7] "Syntax is not the same as, nor by itself is it sufficient for, semantics", [Searle(1992)].

ten, not spoken), has no English, and, moreover, apart from seeming to be able to read and write, is deaf, dumb, blind, and paralyzed.

Why, exactly, should we conclude that Hao doesn't understand the Chinese that he appears to be reading and writing ("automatically", as it were)?"

Haugeland suggests Searle's internalisation response equivocates on the use of the word 'in' and subsequently deduces that Searle is not entitled to assert that, 'were Hao to understand Chinese, so would [Searle]'; hence Haugeland claims Searle's internalisation response to the System reply is simply not logically sound.

4.1 Phenomenal aspects of understanding

To unpack Haugeland's claim, let us compare the responses of the two systems - Hao and Searle - in the case where Searle first listens to a joke in Chinese and then in English. In the former case, although Searle may make the right linguistic responses in Chinese, he will never 'get the joke' and 'feel the laughter' because he, John Searle, still doesn't really understand a word of Chinese; whereas in the latter case he may well 'get the joke', find it funny and laugh because he really does understand English. In other words, the behaviour will not entail understanding, but the understanding will entail the appropriate behaviour.

Similarly, for all a small child may laugh at a sequence of adult jokes, she will not *feel* the laughter appropriately and hence will not really *understand* the jokes. In other words, there is a fundamental 'difference in kind' (an *ontological* distinction) between these two cases. This is perhaps not so surprising as in the former case *ex hypothesi* as Searle-as-Hao is merely carrying out ungrounded, uninterpreted symbol manipulations; whereas in the latter case Searle's command of his native tongue is grounded by consciousness of his body and his interactions with the world and society.

In the absence of any linguistically-grounding 'conscious sensations of laughter' accompanying Searle's execution of the Hao program, we doubt that by any *normal use* of the word 'understand' anyone can, legitimately, claim Searle-as-Hao understands the Chinese story anymore than the young child 'understands' adult jokes; demonstrably [in this case] mere outward behaviour alone is not sufficient to tease apart the two situations.

Hence we suggest that there is a fundamental *ontological* (contra epistemological) distinction between Searle-as-Hao and Searle-as-native-English-speaker; a difference that cannot be teased apart by mere observation of external behaviour alone and that, following Strawson[8], this *conscious* difference is central to the notion of what it really means to 'understand' and to 'think'.

[8] A closely allied position is also endorsed by Horgan and Tienson [Horgan and Tienson(2002)].

But if the CRA is correct and the mere computer simulation of a neural network is not sufficient for understanding, perhaps the addition of a biological neural network would overcome the CRA? We now argue that recent technological advances, which contribute to blurring the divide between biological and artificial systems, may serve as a vehicle to push this examination further. In particular, we focus on so-called animats [Franklin(1997), Wilson(1985)], autonomous robots that are controlled by biological neural tissue, and what may be described as remote-controlled rodents, living animals endowed with augmented abilities provided by artificial controllers. These two chimeras can be seen as the two sides of the same coin and, we argue come a step closer to the physical realisation of the well known "brain-in-a-vat" thought experiment, cousin to the CRA. If correct, our position reinforces Tom Ziemke's distinction between strong and weak embodiment [Ziemke(2001)] and suggests that the former is fully necessary for understanding.

5 Robots and animats

Recently, one of the co-authors led a team at the University of Reading that successfully developed an autonomous robot controlled by cultured living neural cells [Warwick et al(2010a)Warwick, Nasuto, Becerra, and Whalley, ?]. The "brain" of the system consisted of a cultured network of thousands of neurons, sliced from the cortical tissue of foetal rats, and grown on an array of electrodes that permits both recording and electrical stimulation. As a result of the procedure, the connections between the neurons are lost, but within a short period of time, new connections spontaneously form, and neurons start engaging in communication. The activity grows over the subsequent weeks into bursts of activity that spread over the entire culture until maturation (about 1 month after seeding). The resulting activity was then used to control the actuators of a small wheeled robot and, closing the loop of the system, the signal registered by the robot's sensors was being fed back to the cultured neurons in the form of brief electrical impulses. This platform demonstrated simple obstacle avoidance behaviours[9], analogous to a simple Braitenberg vehicle [Braitenberg(1984)], and the *a posteriori* analysis of the cultures showed functional connectivity, as well computational and biophysical properties similar to that of intact brains.

5.1 Programming rodents

In recent years, successes in implant technology gave rise to functional hybrid systems integrating artifacts with the nervous systems of living organisms. Efforts in this direction are motivated by the creation of prostheses, e.g. cochlear [Blake(2000)] or retinal [Fornos et al(2011)Fornos, Sommerhalder, and Pelizzone, Zrenner(2002)] implants, and are now moving beyond augmenting sensory

[9] See www.youtube.com/watch?v=1-0eZytv6Qk.

modalities towards interfacing directly with the brain through deep brain stimulation. This technique involves implanting tiny electrodes in nuclei of the brain, permitting the recording and stimulation of local neurons. It is an approved clinical technique for the treatment of many neurological disorders in humans [Fins(2004), Kringelbach et al(2007)Kringelbach, Jenkinson, Owen, and Aziz]. Before reaching this stage, however, extensive testing has to be performed on seemingly simpler brains, like that of rodents. Recently, implant technology has made huge advances in moving from simple passive electrodes for stimulation to implanting whole electronic circuits capable of performing complex functions. Berger et al., for instance, successfully demonstrated implants that replaced a rat's hippocampus during a spatial memory task: when the device was inactivated, the animal failed the behavioural task; its performance was restored when the device was switched back on [Berger et al(2011)Berger, Hampson, Song, Goonawardena, Marmarelis, and
Another example comes from John Chapin's group, whose implant coupled reward and sensory processing areas in an operant conditioning procedure to train the rat to respond behaviourally to particular tactile stimulations [Talwar et al(2002)Talwar, Xu, Hawley, Weiss, Moxon, and Chapin]. Upon several days of training, the 'programmed rodent' was able to follow commands, henceforth behaving similar to a remote-controlled animal[10]. See also Gradinaru et al [Gradinaru et al(2007)Gradinaru, Thompson, Zhang, Mogri, Kay, Schneider, and Deisseroth], who used optogenetic techniques to stimulate neurons selectively, inducing motor behaviour without requiring conditioning.

5.2 From an "intuition pump" to the physical realisation of thought experiments

Does our animat, which successfully avoids obstacles, genuinely understand it is facing a wall? Can the remote-controlled mouse, which turns left in infinite loops as long as the device is switched on, be said to understand genuinely what it is doing, and why? How about the remote-controlled rat that blindly follows motor commands; does it understand the input it receives?

The situations we described, we posit, push Searle's CRA a little bit further, permitting the philosophical exploration of the fifth line of reply to the CRA, that of the robot endowed with a brain much alike to a biological brain. In these situations, both the animats and what we called remote-controlled rodents experiments assume some degrees of embodiment and relatedness to the workings of biological brains. A systems view would thus legitimately raise the question, "to what degree might these chimeras understand their personal predicament?"

Cosmelli and Thompson already paved the way for this line of thought in an attempt to formulate a response to the "brain-in-a-vat" thought experiment, a cousin to the CRA, about consciousness [Cosmelli and Thompson(2011)]. In this experiment, the reader is invited to imagine a brain floating "in a life-sustaining vat of liquid nutrients" and connected to "a supercomputer that

[10] See www.youtube.com/watch?v=D5u2IWFNFDE.

would stimulate it with electrical impulses exactly like those it normally receives when embodied" (p. 361). Cosmelli and Thompson use this thought experiment to explore the role of the body in the definition of consciousness. Notably, they pose that a functional body is required to support consciousness, and that such a body needs to be "a self-regulating system comprising its own internal, homeodynamic processes and capable of sensorimotor coupling with the outside world" (p. 363), a conception at the heart of the enactive approach to cognitive science, and conclude that "consciousness is a function of life-regulation processes involving dense couplings between neuronal and extraneuronal systems, rather than a function of neural systems alone" (p. 379).

We argue that, even though their interest in this thought experiment lies in the defining features of consciousness, their argument might as well apply to intentionality. In fact, as an analogue of Cosmelli and Thompson's "brain-in-a-vat", we suggest that the impoverished notion of a "body" that serves the animat equally offers no hope for anything more than mere sensorimotor coupling to arise.

The lack of proper embodiment is, however, only part of the problem, as demonstrated by the remote-controlled rodent experiments. In these cases, the chimera is constituted by a fully functional, living body that is endowed with an artificial device that transmits electrical signals to the brain, to which it responds. The crux of this paper is that behaviorally, the remote-controlled animals seem to lose something more than "just" free-will and volition.

This is because there is no element of the animal's intrinsic makeup that would cause it to behave of its own accord in a way similar to that imposed onto it by the experimenter, and hence it is extremely unlikely that it would ever acquire understanding of such externally imposed behaviours[11]. This is in spite of the fact that, in contrast to an animat, this implanted rodent is obviously a case of perfect embodiment, though the experimenter artificially manipulates the rodent's tissues. These manipulations produce sensory-motor couplings that result in the rodent experiencing the world consistent with the induced actions.

The fundamental flaw is that these induced couplings would not be the effect of the intrinsic nervous system's constraints (metabolic or otherwise) at any level. To the contrary, they are the cause of metabolic demands; demands that are incurred through the experimenter's manipulations. Since the experimenter drives the sensorimotor couplings in an arbitrary way (from the perspective of metabolic needs of animal or its cellular constituents), the causal relationship between the bodily milieu and the motor actions and sensory readings would be disrupted. However, according to the enactive approach, only the right type and directionality of such couplings can ultimately lead to understanding and intentionality.

[11] Unless the imposed behaviours happened to exactly synchronise with natural behaviours appropriate to the rodent, given all of its bodily needs and desires, at that point in time.

The above argument is an elaboration on Cosmelli and Thompson's "brain-in-a-vat" critique. A recently proposed synthesis of computational paradigms offered by the authors [Spencer et al(2013)Spencer, Tanay, Roesch, Bishop, and Nasuto] puts forward an alternative but complementary argument which also ties in with the fundamental arguments by Howard Pattee deriving from the cybernetic tradition on the nature of symbols in biological systems [Pattee(1995)] which was also extended towards brain [Cariani(2001)] and higher cognition [Rączaszek-Leonardi(2012)]. It appears that the most encompassing conception of computation [Spencer et al(2013)Spencer, Tanay, Roesch, Bishop, and Nasuto] performed by any system rests on the existence of internal dynamics that is capable of generating a flow in the system state space. The most basic implication of this is that it affords the existence of different system configurations or states; without such potential there can be no computation of any form (continuous or discrete, symbolic or distributed). However, dynamics are not sufficient to instantiate computations, since computations require constraints. It is the constraints that enable the algorithm to 'shape' the flow, and an instantiation of a specific trajectory starting from a specific initial condition corresponds to an execution of a programme. The nature of the state space and that of constraints will be specific for the different computational paradigms but they all share in common that the constraints are externally imposed and are used by the programmer so that the dynamical flow can be meaningfully interpreted. Thus, the constraints do not have an intrinsic meaning in computations, the meaning is imposed on them, and only via their skilful use can the resultant dynamics lead to meaningful results. We posit that, in light of the above, the manipulations of a neural culture in an animat or even of the brain in the remote-rat are nothing else but a form of imposing such external constraints on otherwise innate (neural) dynamics. Moreover, as such constraints are not constructed to fulfill the properties required of so-called "replicable constraints" [Rączaszek-Leonardi(2012)], they must remain merely formal manipulations. Whether at the neuronal or cognitive level, replicable constraints acquire meaning through some selective process rendering them functional at that level. Thus, both animats and remote rats are merely other examples of formal computational systems, albeit 'implemented' in hybrid (animat) or biological (remote rat) substrate.

Therefore, we suggest that, even though the remote rat still possesses a fully functional body and, arguably, a functioning brain, the fact that it receives alien commands does not warrant a genuine understanding of what is going on. In other words, the animal's brain receives foreign input that, at best, may resemble drug-induced decontextualised hallucinations – mere uninterpreted symbols/squiggles and squoggles – which, we argue, would remain meaningless despite how accustomed the animal may become to this new mode of owner-less functioning. The situation is analogous to the alien hand syndrome - where in such patients, for example, their arm seemingly performs actions not of their volition or under their control (in fact, often against their will). Such patients do not accrue any meaning of why their arm acted in this way, albeit they can see (and hence comprehend) the actions in the same way as any other

observer; in this sense they are 'external observers' of their own limb(s) zombie movements.

It is thus clear that genuine intentional states, in the process of understanding the world, require both a fully functional brain and a fully functional body; deporting the question of the requirements for genuine understanding to the defining features of the process whereby the brain and body interact with the world.

Referring back to our computationalist framework [Spencer et al(2013)Spencer, Tanay, Roesch, Bishop, and Nasuto], that means that even if one were to describe cognitive agents in computational terms one is forced to conceive the closest computational analogy of a very special kind. For these are the systems that are not only, following Pattee [Pattee(1995)] and Rączaszek-Leonardi [Rączaszek-Leonardi(2012)], capable of supporting replicable constraints at both biological and cognitive level but which are effectively capable of setting such constraints for themselves, in order to modify their own innate dynamics supporting the 'computation'. It is this self-referential nature of cognitive systems that bootstraps the meaning of replicable constraints/symbols to the meaning of the computation 'executing' by the innate dynamics constraint in this way. It should be also clear from our overview of a wide range of computational paradigms discussed in theoretical computing [Spencer et al(2013)Spencer, Tanay, Roesch, Bishop, and Nasuto] that none of them comes even close to such a system, all being some forms of formal, externally driven symbol manipulation.

Without this, neither animats, nor the remote-controlled rodents experiments can escape Searle's CRA. This argument lends support to discussions of the properties grounding the agent-environment system. Fröese and Ziemke, for instance, discuss the foundational role of constitutive autonomy and adaptivity [Froese and Ziemke(2009)] for agency and sense making, and their consequences for the design of embodied AI.

6 Summary and conclusion

In this paper, we based a philosophical examination of the requirements for genuine understanding and intentionality on extensions to John Searle's Chinese Room Argument against strong AI. Our deployment of Searle's "intuition pump" to recent advances in robotics and neuroscience shows it continues to have force against the most recent developments in robotics and bio-machine hybrids. Specifically, we examined how two new scenarios fare in light of the CRA's typical replies from the AI community. We focused on so-called "animats", autonomous robots that are controlled by biological neural tissue, and what may be described as "remote-controlled rodents", living animals endowed with augmented abilities provided by artificial controllers. These two chimeras can be seen as the two sides of the same coin, and herein we have demonstrated that neither of these systems can be said to exhibit any genuine understanding of the world.

In addition, we argue that current efforts in cognitive robotics, to endow robots with abilities to represent the world and reason about it, are limited. In line with the rise of enactive cognitive science - which proposes an enlarged perspective that includes the closed-loop interactions of a life-regulated body-brain dynamical system with an evolving world - we deem inappropriate cognitivism and its concomitant computational theory of mind, and instead emphasise the role of foundational processes such as autonomy, exploration, autopoiesis and social embedded-ness, in giving rise to a genuine understanding of our lived world.

Acknowledgements We would like to thank Dr. Tom Fröese for comments which helped improve this paper. Please note that, in the context of the '*Computing, Philosophy and the Question of Bio-Machine Hybrids: 5th AISB Symposium on Computing and Philosophy*' [part of the 2012 Turing centenary AISB/IACAP Joint World Congress], elements of this work re-visit arguments first raised at the 2011 PT-AI conference, Thessaloniki. C.f. Vincent C. Müller (ed.), (2012), *Theory and Philosophy of Artificial Intelligence*, (SAPERE; Berlin: Springer).

References

[Berger et al(2011)Berger, Hampson, Song, Goonawardena, Marmarelis, and Deadwyler] Berger TW, Hampson RE, Song D, Goonawardena A, Marmarelis VZ, Deadwyler SA (2011) A cortical neural prosthesis for restoring and enhancing memory. Journal of Neural Engineering 8(4):046,017

[Bishop(2004)] Bishop JM (2004) A view inside the chinese room. Philosopher 28(4):47–51

[Bishop(2009)] Bishop JM (2009) A cognitive computation fallacy? cognition, computations and panpsychism. Cognitive Computation 1(3):221–233

[Blake(2000)] Blake W (2000) Cochlear Implants: Principles and Practices. Lippincott Williams & Wilkins, Philadelphia

[Boden(1988)] Boden M (1988) Escaping from the chinese room. In: Boden M (ed) The philosophy of Artificial Intelligence, Oxford University Press, Oxford, UK, pp 89–105

[Braitenberg(1984)] Braitenberg V (1984) Vehicles: experiments in synthetic psychology. MIT Press, Cambridge, MA, USA

[Bringsjord(2002)] Bringsjord R S & Noel (2002) Real robots and the missing thought-experiment. In: Preston J, Bishop JM (eds) Views into the Chinese Room:: new essays on Searle and Artificial Intelligence, Oxford University Press, Oxford, pp 360–379

[Cariani(2001)] Cariani P (2001) Symbols and dynamics in the brain. Bio Systems 60(1-3):59–83

[Cassirer(1944)] Cassirer E (1944) An Essay on Man. Yale University Press, New Haven

[Cole(2009)] Cole D (2009) The chinese room argument. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, winter 2009 edn, Stanford University

[Cosmelli and Thompson(2011)] Cosmelli D, Thompson E (2011) Embodiment or envatment? reflections on the bodily basis of consciousness. In: Stewart J, Gapenne O, , di Paolo E (eds) Enaction: Towards a New Paradigm for Cognitive Science, MIT Press

[Dennett(1991)] Dennett D (1991) Consciousness Explained. The Penguin Press, Allen, Lane

[Fins(2004)] Fins JJ (2004) Deep brain stimulation. In: Post SG (ed) Encyclopedia of Bioethics, vol 2, 3rd edn, MacMillan Reference, New York, pp 629–634

[Fodor(1975)] Fodor JA (1975) The Language of Thought. Harvard University Press

[Fodor(1987)] Fodor JA (1987) Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT Press

[Fornos et al(2011)Fornos, Sommerhalder, and Pelizzone] Fornos A, Sommerhalder J, Pelizzone M (2011) Reading with a simulated 60-channel implant. Frontiers in Neuroscience 5:57

[Franklin(1997)] Franklin S (1997) Artificial Minds. MIT Press
[Freeman(2003)] Freeman A (2003) Output still not really convinced. The Times Higher
[Freeman(2004)] Freeman A (2004) The chinese room comes of age: a review of preston & bishop. Journal of Consciousness Studies 11(5-6):156–158
[Froese and Ziemke(2009)] Froese T, Ziemke T (2009) Enactive artificial intelligence: Investigating the systemic organization of life and mind. Artificial Intelligence 173(3–4):466–500
[Garvey(2003)] Garvey J (2003) A room with a view? The Philosophers Magazine 23:61–61
[Gradinaru et al(2007)Gradinaru, Thompson, Zhang, Mogri, Kay, Schneider, and Deisseroth] Gradinaru V, Thompson KR, Zhang F, Mogri M, Kay K, Schneider MB, Deisseroth K (2007) Targeting and readout strategies for fast optical neural control in vitro and in vivo. J Neurosci 26:27(52):14,231–14,238
[Harré and Wang(1999)] Harré R, Wang H (1999) Setting up a real 'chinese room': an empirical replication of a famous thought experiment. Journal of Experimental & Theoretical Artificial Intelligence 11(2):153–154
[Haugland(2002)] Haugland J (2002) Syntax, semantics, physics. In: Preston J, M BJ (eds) Views into the Chinese room: new essays on Searle and Artificial Intelligence, Clarendon Press, Oxford, UK
[Horgan and Tienson(2002)] Horgan T, Tienson J (2002) The intentionality of phenomenology and the phenomenology of intentionality. In: Chalmers D (ed) Philosophy of Mind: Classical and Contemporary Readings, Oxford University Press, Oxford, UK, pp 520–33
[J.(2004)] J R (2004) Review of preston j and bishop m, editors. views into the chinese room: New essays on searle and artificial intelligence. Philosophical Books 45(2):162–167
[Kringelbach et al(2007)Kringelbach, Jenkinson, Owen, and Aziz] Kringelbach ML, Jenkinson N, Owen SLF, Aziz TZ (2007) Translational principles of deep brain stimulation. Nature Reviews Neuroscience 8:623–635
[Newell(1976)] Newell HA A & Simon (1976) Computer science as empirical inquiry: symbols and search. Communications of the ACM 19(3):113–126
[Overill(2004)] Overill J (2004) Views into the chinese room: New essays on searle and artificial intelligence. Journal of Logic and Computation 14(2):325–326
[Pattee(1995)] Pattee HH (1995) Evolving self-reference: matter, symbols, and semantic closure. Communication and cognition-artificial intelligence 12(1-2):9–27
[Pfeifer(2001)] Pfeifer C R Scheier (ed) (2001) Understanding Intelligence. MIT Press
[Preston and Bishop(2002)] Preston J, Bishop JM (eds) (2002) Views into the Chinese room: new essays on Searle and Artificial Intelligence. Oxford University Press, Oxford, UK
[Putnam(1988)] Putnam H (1988) Representation and Reality. MIT Press
[Rapaport(2006)] Rapaport WJ (2006) Review of preston j & bishop m, editors. views into the chinese room: New essays on searle and artificial intelligence. Australian Journal of Philosophy 94(1):129–145
[Rey(2002)] Rey G (2002) Searle's misunderstanding of functionalism and strong ai. In: Preston J, Bishop JM (eds) Views into the Chinese Room:: new essays on Searle and Artificial Intelligence, Oxford University Press, Oxford, pp 360–379
[Rączaszek-Leonardi(2012)] Rączaszek-Leonardi J (2012) Language as a system of replicable constraints. LAWS, LANGUAGE and LIFE pp 1–34
[Roesch(in press)] Roesch EB (in press) A critical review of classical computational approaches to cognitive robotics: Case study for theories of cognition. In: Radman Z (ed) The Hand: An Organ Of The Mind, MIT Press, Cambridge, USA
[Schank and Abelson(1977)] Schank RC, Abelson RP (1977) Scripts, plans, goals and understanding: An inquiry into human knowledge structures. Erlbaum, Hilsdale NJ
[Searle(1980)] Searle J (1980) Minds, brains, and programs. Behavioral and Brain Sciences 3:417–457
[Searle(1992)] Searle J (1992) The Rediscovery of the Mind. Bradford Books
[Spencer et al(2013)Spencer, Tanay, Roesch, Bishop, and Nasuto] Spencer MC, Tanay T, Roesch EB, Bishop JM, Nasuto SJ (2013) Abstract platforms of computation. In: AISB 2013
[Sprevak(2005)] Sprevak MD (2005) The chinese carnival. Studies in the History & Philosophy of Science 36:203–209

[Talwar et al(2002)Talwar, Xu, Hawley, Weiss, Moxon, and Chapin] Talwar SK, Xu S, Hawley E, Weiss S, Moxon K, Chapin J (2002) Rat navigation guided by remote control. Nature 417:37–38

[Warwick et al(2010a)Warwick, Nasuto, Becerra, and Whalley] Warwick K, Nasuto SJ, Becerra VM, Whalley BJ (2010a) Experiments with an in-vitro robot brain. In: Cai Y (ed) Instinctive Computing, Lecture Notes in Artificial Intelligence, vol 5987, Springer

[Warwick et al(2010b)Warwick, Xydas, Nasuto, Becerra, Hammond, Downes, Marshall, and J] Warwick K, Xydas D, Nasuto SJ, Becerra VM, Hammond MW, Downes JH, Marshall S, J WB (2010b) Controlling a mobile robot with a biological brain. Defence Science Journal 60(1):5–14

[Waskan(2005)] Waskan JA (2005) Review of preston j and bishop m, editors. views into the chinese room: New essays on searle and artificial intelligence. Philosophical Review 114(2):277–282

[Wilson(1985)] Wilson SW (1985) Knowledge growth in an artificial animal. In: Grefenstette JJ (ed) First International Conference on Genetic Algorithms and Their Applications, Lawrence Erlbaum Associates, Hillsdale, NJ, pp 16–23

[Wittgenstein(1958)] Wittgenstein L (1958) Philosophical Investigations. Blackwell, Oxford, UK

[Ziemke(2001)] Ziemke T (2001) Are robots embodied? In: Lund University Cognitive Studies, pp 75–83

[Zrenner(2002)] Zrenner E (2002) Will retinal implants restore vision? Science 295:1022–1025