Symposium on Philosophy after AI: Mind, Language and Action

In conjunction with the 2018 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2018)

6th April 2018

Principles of Robot Autonomy

Steve Battle [1]

Abstract. This paper is an exploration of the theory of *autonomy* as set out by Francisco J. Varela in his groundbreaking book, “Principles of Biological Autonomy.” Varela argues that autonomy is an expression of a pervasive circularity, or indeed complete organisational closure. This closure is seen as a necessary condition for recursively maintaining a system’s identity, or *self*. This paper applies these ideas to robotics in the quest for a theoretical understanding of robot autonomy as distinct from artificial-life and artificial-intelligence. Using directed-graph based descriptions of behaviour, we explore centrality measures of their topological properties that are consonant with autonomy.

1 INTRODUCTION

When we think about deep questions such as “What is life?”, we may think back to the checklist of characteristic traits we learned in school, including metabolism, growth, reproduction, etc. In contrast, a systems, or cybernetic view of ‘life’ can provide a more satisfying account in terms of *autopoiesis* [13], whereby a living system is understood as a network of processes of continuous, physical self-creation. From Maturana and Varela’s *Autopoiesis and Cognition*, “An autopoietic machine continuously generates and specifies its own organisation through its operation as a system of production of its own components.” The scientific investigation of life often emphasises the reproductive capacity of life, wheres autopoiesis focuses on the organisational and structural coherence of the individual organism. The organisation of a system is the network of physical processes of production; in a living organism this would include the expression of genes as proteins and other products that play a key role in the maintenance of life. The second key aspect of autopoiesis is the way this organisation maintains the physical structure of the system; the unity of the organism. For example, an autocatalytic chemical reaction is not autopoietic because while it may have the necessary organisation, it lacks the spatial, cellular structure that characterises the unity of an organism [13, p. 94].

The way autopoietic processes turn in on themselves is known as *organisational closure*. It constitutes a homeostatic system [1] whose function is to maintain itself through the synthesis of its own components, and the maintenance of the structure that houses it. The autopoietic system is the physical realisation of an attractor that tends towards some dynamically stable behaviour. Even if disturbed it can return to stability, unless it is pushed too far from the orbit of the attractor and it disintegrates. Autopoiesis, or self-production, is what it means to be alive.

Autopoiesis provides a truly inspiring view of living-machines. Maturana states that, “Living systems are cognitive systems, and living as a process is a process of cognition” [13, p. 13]. However, “Cognition is a biological phenomenon *and can only be understood as such*” [13, p. 7]. This is a crushing conclusion for engineers seeking to apply these lessons to cognitive robots. The science of synthetic artificial-life is only now emerging, but what hope is there for those of us versed in the ways of non-living computers and robotics?

2 AUTONOMY AND THE SELF

Living systems are fundamentally autonomous by nature. The precursor to ‘Autopoiesis’ was Varela’s [16], “Principles of Biological Autonomy.” While following the same general scheme as ‘Autopoiesis’, this book is aimed squarely at describing the nature of *autonomy*. The focus shifts away from the physical processes of production, to processes for the maintenance of *identity*. While all autopoietic systems are autonomous by definition, the family of autonomous systems may be expanded to admit neurological and computational processes, and even conversational interactions; machines both living and non-living. Varela argues [16, p. 55] that these processes “are related as a network, so that they recursively depend on each other in the generation and realisation of the processes themselves.” The *identity* of the system can be seen as the coherent organisation of this dynamic unity. If the process is disrupted, then this unity disappears and the system suffers loss of identity. For an autopoietic system a similar disruption would literally mean loss of life.

As with autopoiesis, organisational closure is key to understanding autonomy. Whilst organisational closure is inward looking, an autonomous system *may* be disturbed, or perturbed, by external events, and may, in turn, push back on the world as it compensates for these disturbances. However, one counter-intuitive consequence of organisational closure is that both autopoietic and autonomous machines have *no inputs or outputs*. We would think of an input or output as a variable (sensor or actuator) that is open to the system’s environment, but this would have the effect of breaking any cycle it belonged to. We should think of these variables as operating primarily within closed cycles of influence. This indirect engagement with the environment through perturbation, or deformation, is known as *structural coupling*, to distinguish it from simple input/output. The *domain of interactions* of an autonomous system is the domain of all deformations the system may undergo without loss of identity. A simple robot may be structurally coupled with features of its environment; reacting to their presence though the behaviour it exhibits to compensate for the original perturbation.

We saw above that no information passes into or out of the autonomous system, as the organisationally closed network makes no references to external symbols, being defined only in terms of its internal variables. Such information can really only be understood in relation to the observer. In making sense of the interaction with an autonomous system, the observer may find pattern and meaning in the behaviour in relation to its environment. In doing so, the ob-

[1] Computer Science and Creative Technologies, University of the West of England, Bristol, UK, email: steve.battle@uwe.ac.uk

server describes the *cognitive domain* of the system in relation to its environment. Despite this acknowledgement of the role of the observer, we must not lose sight of the fact that the observer independent maintenance of identity is the central function of the organisationally closed, autonomous system.

We may contrast autonomous systems with *allonomous* systems that are open to the environment, having inputs and outputs, and open to external control; in effect something like a computer. Computers are open to external instruction - programming - which means that their entire *cognitive domain*, the domain of interactions possible with their environment, is externally imposed. Cybernetic thinking about autonomous systems rejects the "gestalt of the computer" where the computer is regarded not only as a tool, but as a way of thinking about the world [14]. As AI researchers we are constantly aware that the symbols we impose on the computer in the form of code demonstrate the power of the computer as an *allonomous* tool, but expose a limitation in that the meaning of these symbols is no longer the *autonomous* product of the system itself, for its own ends.

Complex dynamic systems are often described using systems of simultaneous equations, but Varela explored new ways in which symbolic approaches could be used to the same effect. The operational stability of a system, numeric or symbolic, can be analysed by finding its *eigenvalues*, or put more generally its *eigenbehaviours*. These are analogous to the resonant frequencies of a system, the modes of oscillation that represent recursive cycles of self-realisation.

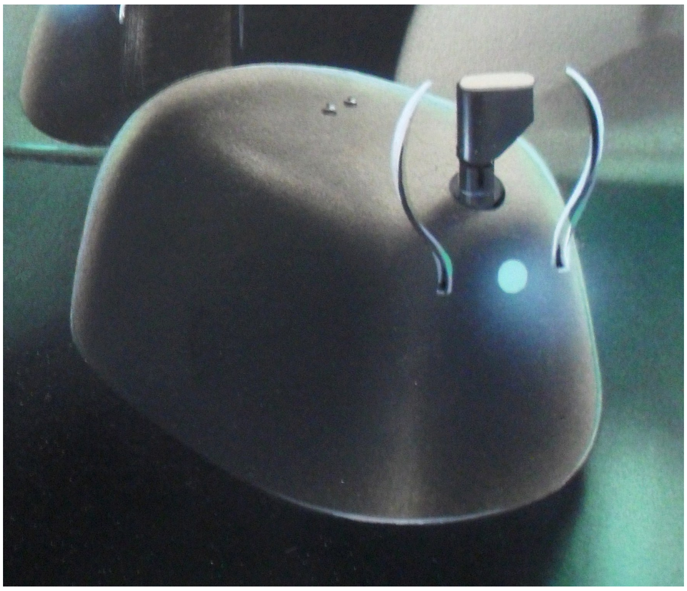

Figure 1. ELSIE: Electro-mechanical robot, Light Sensitive with Internal and External stability. Her scanning turret contains a single photo-cell, and a tilt of her shell detects contact with obstacles.

3 ROBOT AUTONOMY

Varela's definition of autonomy provides us with a tool for understanding such systems as a superset of living-machines. Varela applies the idea to the nervous and immune systems, and it is also possible to apply the theory to autonomous robots. Autonomy is distinct from cognition, learning and intelligence, so can be used to analyse the very simplest examples of autonomous robot in our menagerie. The aim of this analysis is to understand the organisationally closed dynamic processes that maintain a robot's identity as an autonomous system.

The first truly autonomous robot was, arguably, William Grey Walter's 'ELSIE' [8] as illustrated in figure 1 by artist Robin Day for the Festival of Britain in 1951 [11]. Built in Bristol in 1948, ELSIE is an Electro-mechanical robot, Light Sensitive with Internal and External stability. She is autonomous in the sense that she can 'explore' her environment, seeking light sources but keeping a respectful distance from them (external stability) being both positively and negatively phototropic. If she encounters an obstruction, the displacement of her shell activates a 'trembler' switch causing an oscillatory "push and turn" behaviour that will free her of the obstacle. In addition, when her batteries run low, she loses her negative phototropism and will move towards a conveniently illuminated battery charger (internal stability). While recharging, her motors are disconnected, but when the charging current drops sufficiently they reconnect and she pushes back from the charger as a consequence of negative phototropism.

4 TELEOLOGICAL DESCRIPTION

The role of the observer is key in Varela's theory of autonomy, as he makes a strong case for distinguishing the cognitive domain, open to the observer, from the operational or causal processes underlying the phenomenology of the system. We can observe behaviours such as "keeping a respectful distance", "freeing her of the obstacle" and "moving towards a battery charger." These *teleological* explanations, expressed in terms of purposes and aims, do not belong to the system itself. Purpose belongs to the domain of observation. Such explanations are shorthand for patterns of behaviour in the context of an environment. However, this doesn't fundamentally invalidate this kind of *symbolic* explanation; Varela supports what he calls *descriptive complementarity* where no one explanation should be seen as more fundamental than the other. Varela introduces his *star* notation to describe this complementarity:

$$* = \text{symbolic}/\text{operational}$$

This is understood to mean that not only do symbolic and operational descriptions differ, but furthermore that they are different levels of description. The stroke, '/', is not just a division between the two viewpoints, but indicates the existence of a method for getting from one to the other; an unfolding of the operational model produces a symbolic description, and methods for capturing an operational model based on the evidence of the symbolic domain. It also expresses the tension between holist and reductionist viewpoints. Instead of seeing these as incompatible opposites, we might view them as complementary and mutually supporting viewpoints. Thus there is a place for high-level teleological descriptions, complemented by lower level causal models. While these causal, operational models enable prediction, symbolic explanations have a pedagogical function enabling the system to be discussed and understood.

A symbolic (teleonic in Varela's terms) explanation focuses on patterns of observed behaviour, described in a way that aids sense-making. In particular it may focus on the interaction of the system with its environment. Grey Walter's notes [9] on the operation of the Machina Speculatrix (ELSIE) identify a number of symbolic patterns of behaviour. One of these is selected from moment to moment, depending on the sensed light level, contact with the shell, and battery level. An additional behaviour R has been added to represent the state the robot is in when it is recharging, with motors stopped.

- *Pattern E* - Exploration
- *Pattern P* - Positive phototropic response
- *Pattern N* - Negative phototropism
- *Pattern O* - Obstacle avoidance
- *Pattern R* - Recharging

This set of patterns can be thought of as the cognitive domain of the robot, as each one is triggered by, and compensates for, a perturbation caused by the environment. Furthermore, their descriptions are couched in teleonic, observer relative terms. Each pattern describes a particular *state* of the system representing a particular electro-mechanical configuration. Larger patterns only emerge when the system is placed within an environment containing obstacles and light sources. These larger patterns can be seen in strings of plausible behaviours, based on the symbolic codes above. To be clear, these behaviours are the result of thought experiments so the frequency of occurrence of any symbol or sequence of symbols should be disregarded. The first string shows a *progressive orientation* towards a light source. With E active in low-light levels, P is the initial sighting of the light source, and as it gets closer to the source, the negative phototropic response, N, kicks in. The alternation between E, P and eventually N is caused by the rotating scanning turret of the robot containing a single photo-cell 'eye'. Assume that the light level detected varies continuously so there can be no sudden jump from E to N or vice versa. In the 2$^{\text{nd}}$ trace, an obstacle is encountered multiple times which can occur at any point in the scanning cycle, so can follow or precede any of E, P, N. The recharging state, R, can only be achieved through the positive response of moving towards the illuminated charger (with suppressed negative phototropism), and the fully charged robot exits with the opposite negative phototropism, N. Only state changes are recorded so that no symbol follows itself.

$$\begin{array}{l}
EPEPEPNPEPNPE\ldots \\
EPEOEPONOEOPEPEPE\ldots \\
EPEPOPNPEPRNPE\ldots \\
\qquad\qquad\vdots
\end{array}$$

We can think of the behavioural traces above, as a single tree with strings sharing a common root becoming independent branches off that root. The three examples share a common root "EPE." At a minimum they would share the empty string. The tree is potentially infinitely deep.

5 OPERATIONAL DESCRIPTION

An operational description of the robot must generate the observed behaviour and should be expressed in a form that is organisationally closed. This is achieved using a *directed* graph, G, as shown in figure 2. While ELSIE is a very simple machine, state emerges from the angular position of the rotating turret, her physical location relative light sources and obstacles in the environment, and current battery level. State is therefore not just internal to the robot, but a property of the robot within its environment. In the treatment below, ELSIE's observable behaviour is captured as a state-machine. Each state is associated with a symbol from the cognitive domain, just one of which may be emitted in each state. These can be only incidental to the operational description as they have no internal, causal role.

State-machines are used here because they provide us with one of the simplest models available for exploring autonomous behaviour, and in particular the simple behaviour of ELSIE. The only real sense in which they help us escape the binds of the "computer gestalt" is to lower our expectations about the symbolic power of the underlying operational model. In their simplicity they enable us to define a self-contained, organisationally closed model. The state-machine representation is sufficient to capture the regular sequence of symbolic 'observations' we saw above. Such finite automata can generate the regular languages described by Chomsky Type-3 grammars. As we have seen, they should have no symbolic input; the 'observed' behaviours are simply the behaviours of the system embedded within its environment, without further analysis of the nature of the perturbations caused by the environment. Conversely, the system should have no output, so the *emission* matrix is purely for the convenience of the observer.

It is possible to define an operational model of this system as a non-deterministic Finite State Machine with six states. In each state, the machine may emit at most one of the symbols E, P, N, O, R as indicated in the graph. Each of the behavioural traces above can be seen as an *unfolding* of G over time. The graph can be represented compactly as the square adjacency matrix, A, below, with each row and column representing states 1 to 6. The presence of a 1 in row i, column j indicates a directed edge, or transition from state i to state j, while a 0 indicates that no such transition is possible. An incidence matrix records the relationship between a given state in row i, and a possible output symbol in columns labeled E, P, N, O, R, respectively. In this case, each row contains only a single incidence. For example, in both states 1 & 4 the symbol E may be emitted, so column E has 1 in rows 1 & 4.

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad I = \begin{array}{c} \begin{array}{ccccc} E & P & N & O & R \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

The relationship between graph $G = (V, E)$ with $|V| = n$ vertices (nodes) and directed edges (arcs) E, and the adjacency matrix A is defined in equations 1 and 2 as follows. As G is directed (a digraph), the matrix A is not symmetric.

$$A_{ij} = \begin{cases} 1, & \text{if } (i, j) \text{ is a directed edge in } E \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$I_{ij} = \begin{cases} 1, & \text{if } (E, P, N, O, R)_j \text{ may be emitted in state } i \epsilon V \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

6 EIGENCENTRALITY

An eigenvalue denotes the fixed points of a transformation derived by iteration over the operational definition of the system, encompassing the classical notion of stability. Generalising the concept of *eigenvalue*, Varela proposes the name *eigenbehaviour* to refer to the autonomous behaviour of a concrete system such as our robot. The adjacency matrix A is in a form where we can find its eigenvalues and eigenvectors using matrix arithmetic. A scalar λ is an eigenvalue

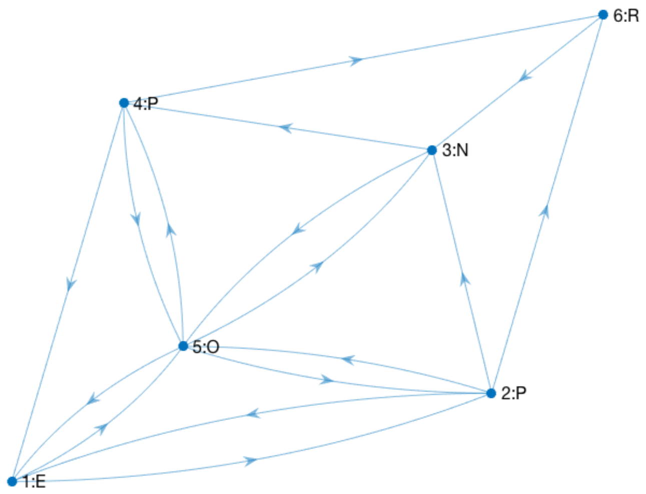

Figure 2. Operational state-machine based description

of A if there is a non-zero vector X such that $AX = \lambda X$. The vector X is called an eigenvector of A corresponding to λ. The eigenvalues λ_i of A are given as a column vector below. The eigenvalue with the greatest magnitude ($\lambda_1 = 2.8111$) is known as the spectral radius.

$$\lambda_i = \begin{pmatrix} 2.8111 \\ -0.1299 \\ -0.1299 \\ -1.5512 \\ -1.0000 \\ 0.0000 \end{pmatrix} \qquad Q_{j=1} = \begin{pmatrix} 0.3819 \\ 0.5026 \\ 0.3390 \\ 0.3819 \\ 0.5711 \\ 0.1206 \end{pmatrix}$$

These eigenvalues are distinct, so the matrix A is diagonalisable and has an *eigendecomposition* into its eigenvalues and eigenvectors. $A = Q\Lambda Q^{-1}$ where Λ is the diagonal matrix whose diagonal elements are the eigenvalues λ_i, Q is a square matrix whose columns are the eigenvectors of A, and Q^{-1} is the inverse matrix of Q.

Centrality measures tell us which nodes are important in a graph. The eigenvector corresponding to the spectral radius is called the principal eigenvector, in this case, found in column 1 of Q. This eigenvector tells us something about the eigenvector centrality, or just *eigencentrality*. This indicates the *importance* of a node based on its out-degree (out-degree eigencentrality). Importance is rated according to the number of outgoing connections a node has, weighted in turn by their importance. The nodes are ranked in order 5,2,1,4,3,6 (with 1 & 4 tied) which can be seen to be broadly in line with their out-degree, and the out-degree of their successors.

If we transpose the adjacency matrix and recompute the eigenvectors then this would be based on the number of incoming connections; also known as *prestige*. The resulting in-degree eigencentrality is more familiarly known as 'page-rank' and is the kind of centrality measure used by internet search engines to rank web-pages. However, this doesn't capture the key characteristic of autonomous systems, namely cyclic, oscillatory patterns of behaviour. An alternative centrality measure must be found.

7 SUBGRAPH CENTRALITY

The number of walks of length k between any two nodes in the graph can be computed by raising the adjacency matrix to the power k, or A^k. A closed walk in a graph is a succession of edges starting and ending at the same node. Subgraph centrality [5] is defined as the weighted sum of the closed walks of length k starting and finishing at a given node. The titular subgraphs are the digons, triangles, squares, etc. that form closed walks within the directed graph. The number of closed walks of length k starting and finishing at node i is $[A^k]_{ii}$. However, the sum of the series of closed walks diverges as k tends to infinity. To achieve convergence a weighting of $1/k!$ is applied, with the effect that short walks have more influence on the centrality of the node than long walks. The subgraph centrality, SC, of A for node i is defined in equation 3 below, where Λ remains a diagonal matrix before and after exponentiation.

$$SC(i) = \sum_{k=0}^{\infty} \frac{[A^k]_{ii}}{k!} = [e^A]_{ii} = [Qe^\Lambda Q^{-1}]_{ii} \tag{3}$$

The subgraph centrality is the diagonal entry of the matrix exponential of the adjacency matrix, e^A [6]. For the eigenvalues and eigenvectors of A, SC is calculated to be as follows.

$$SC = \begin{pmatrix} 3.4448 \\ 3.5052 \\ 2.7774 \\ 2.7170 \\ 5.6360 \\ 1.3266 \end{pmatrix}$$

Centrality allows us to rank nodes according to the topological features that they measure. Given Varela's definition, autonomy is all or nothing and admits of no degree, but we can consider individual nodes' contributions towards autonomy. According to the result, SC, the nodes may be ranked in order 5, 2, 1, 3, 4, 6. The result is little different to the out-degree eigencentrality calculated above, but we see that node 3 has gone up in the ranking because of its involvement in more cyclic walks.

8 CENTRALITY MEASURES COMPARED

Eigencentrality measures diverge from subgraph-centrality with the addition of nodes that are well-connected but do not participate in cyclic walks. Consider adding an additional node to the directed graph representing a failure state where the battery has run flat. It is possible to enter this state directly from all of the other states (1-5) except those that emit symbol R representing recharging (state 6), but there can be no exit from this new sink node (state 7). There is no need to add a new output symbol for this state, as this failure state can be represented in the incidence matrix by a row of 0's, indicating that no symbols are emitted (not shown). Call, A', the adjacency matrix corresponding this extended graph.

$$A' = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The in-degree centrality, or *prestige*, of this graph is computed by decomposing the transpose of matrix A' into its eigenvalues and eigenvectors, $A' = Q'\Lambda'Q'^{-1}$. This time, the maximum eigenvalue is the 2nd element on the diagonal, λ'_i of Λ' where $\lambda'_2 = 2.8111$, so the principal eigenvector is column 2 of Q'. It can be seen that the new sink node has the maximum prestige, corresponding to it having the greatest in-degree (in-degree=5).

$$\lambda'_i = \begin{pmatrix} 0 \\ 2.8111 \\ -1.5512 \\ -1.0000 \\ -0.1299 \\ -0.1299 \\ 0.0000 \end{pmatrix} \qquad Q'_{j=2} = \begin{pmatrix} 0.3586 \\ 0.2860 \\ 0.3314 \\ 0.2764 \\ 0.4455 \\ 0.2001 \\ 0.6040 \end{pmatrix}$$

If we compute the subgraph centrality for this adjacency matrix we obtain the following. The utility of subgraph centrality is that it accounts for the participation of nodes in autonomous cycles. The addition of a sink node creates no extra walks that begin and end at the same node, so the centrality measures, and therefore the ranking of nodes 1 to 5, remain unchanged. The centrality measure of the new node is 1. For a given centrality measure of a node, i, in a directed graph, $SC(i) > 1$ only if there is at least one closed walk that starts and finishes at that node [10]. A measure of 1 indicates that the node does not participate in the autonomous behaviour of the system.

$$SC' = \begin{pmatrix} 3.4448 \\ 3.5052 \\ 2.7774 \\ 2.7170 \\ 5.6360 \\ 1.3266 \\ 1.0000 \end{pmatrix}$$

As the extra sink node is not visited within any cycle it does not form part of the dynamically generated *unity*. The identity of the autonomous system therefore extends only to the subgraph comprising nodes 1 to 6 where $SC'(i) > 1$. This chimes with our intuitions that any sink node terminates the autonomous cycles that are necessary for maintaining identity.

A centrality measure closely related to subgraph centrality is *total communicability*. Where Subgraph centrality is based on the number of closed walks from a node back to itself, *communicability*, as defined in equation 4 below, is based on a weighted sum over all possible walks between pairs of nodes [7].

$$GC_{ij} = \sum_{k=0}^{\infty} \frac{[A^k]_{ij}}{k!} = [e^A]_{ij} \tag{4}$$

The i^{th} row sum of e^A counts the total number of walks between node i and every other node in the graph, including closed walks back to i. As with subgraph centrality, each walk is weighted by $1/k!$ where k is the length of the walk. The i^{th} row sum of e^A is the *total subgraph communicability* of node i [2]. As this is a sum over outgoing edges, the sink node will achieve a low rank. However, if the adjacency matrix is transposed then the sink node becomes a source node with in-degree, 0. Like sink nodes, source nodes do not contribute to the cyclic behaviour that is characteristic of autonomy. The number of closed walks is unaffected, so the subgraph centrality remains unchanged. Unsurprisingly, the source node ($i = 7$) is ranked highest of the i^{th} row sums as it has a high out-degree and there are many walks starting out from this node.

$$GC(A'^T) = \begin{pmatrix} 17.4646 \\ 13.9060 \\ 16.5865 \\ 13.7153 \\ 21.7604 \\ 10.3561 \\ 28.8689 \end{pmatrix}$$

Because source nodes are highly ranked compared with subgraph centrality, we find that total communicability centrality does not chime well with autonomy.

9 NETWORK ROBUSTNESS

Rather than focusing on individual nodes, it is possible to calculate a numerical index characterising the graph as a whole. The so-called Estrada index [3, 4] is defined in equation 5 below, to be the sum of the elements of the vector, SC, representing the subgraph-centrality of each node. This is equivalent to the *trace* (sum of the diagonal) of the adjacency matrix exponential. Again, let $G = (V, E)$, where $|V| = n$, with adjacency matrix A as defined as in equation 1.

$$EE(G) = \sum_{i=0}^{n} e^{\lambda i} = tr(e^A) \tag{5}$$

The robustness of an autonomous system can be seen as the degree of redundancy in the number of closed walks from any node back to itself. If one walk should be unavailable, then another may be taken in its place. The Estrada index grows quickly for large numbers of nodes, so the natural logarithm of the averaged Estrada index may be used as a measure of graph robustness, also known as the *natural connectivity* of the graph [12], defined in equation 6 below.

$$\bar{\lambda} = ln\left(\frac{EE(G)}{n}\right) = ln\left(\frac{tr(e^A)}{n}\right) \tag{6}$$

Following equation 5, the Estrada index of the graph G, $EE(G) = 19.407$, and the corresponding *natural connectivity* as defined in equation 6, $\bar{\lambda}(A) = 1.1739$.

The graph may be perturbed to see the effect of loss of connectivity between nodes [15]. One example of a perturbation that can be applied to graph G is the loss of all transitions to a node. The mask M is used to mask *out* all transitions into node 5, representing the obstacle avoidance behaviour O. First invert the mask by subtracting each element from 1, then take the Hadamard (entry-wise) product to give the perturbed adjacency matrix, $A'' = A \circ (1 - M)$. This is equivalent to placing the robot in an environment without obstacles, or even removing the contact sensor altogether. Either way, this results in an impoverished range of behaviours exhibited by the robot.

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

The robustness of the graph changes monotonically with the removal (or addition) of edges [12, 17]. We therefore expect the perturbed graph adjacency matrix, A'', to have a lower index than that for A.

$$\bar{\lambda}(A'') = \bar{\lambda}(A \circ (1 - M)) = 0.26 < \bar{\lambda}(A) = 1.1739$$

This provides us with a good measure for assessing the robustness of an autonomous system.

10 CONCLUSION

Francisco Varela's ground-breaking work on the principles of autonomy opens the way to applying concepts derived from Maturana and Varela's *autopoiesis*, to non-living autonomous robots. The concept of organisational closure allows us to factor out concerns about inputs and outputs, focusing instead on internal, autonomous processes of identity maintenance. Using state machines to construct operational descriptions of operationally closed systems, centrality measures were identified that reflect properties of the graph topology that support autonomy. Subgraph centrality addresses the cyclic, or oscillatory patterns of behaviour that are the hallmark of autonomy. By this measure we may rank different nodes, or states, in terms of their contribution to the autonomous behaviour and maintenance of identity. In the extreme we can identify source and sink nodes that play no part in this process of self-realisation. Subgraph centrality and the Estrada Index feed into the calculation of a robustness measure for autonomous networks in which we can see the effect of perturbations such as the removal of transitions, that impoverish the range of available behaviours.

11 APPENDIX: MATLAB CODE

```
A = [
0 1 0 0 1 0;
1 0 1 0 1 1;
0 0 0 1 1 0;
1 0 0 0 1 1;
1 1 1 1 0 0;
0 0 1 0 0 0 ];
g=plot(digraph(A))
labelnode(g,[1 2 3 4 5 6],
{'1:E','2:P','3:N','4:P', '5:O','6:R'})
axis off
% Compute eigenvectors/values Q, D
[Q,D] = eig(A)
% Eigenvalues of A
real(eig(A))
% Principal eigenvector
real(Q(1:6,1))
% Compute subgraph centrality SC
SC = Q*diag(exp(diag(D)))*inv(Q)
% Diagonal entry of the subgraph centrality
real(diag(SC))
% A' = A with additional sink node
A1 = [
0 1 0 0 1 0 1;
1 0 1 0 1 1 1;
0 0 0 1 1 0 1;
1 0 0 0 1 1 1;
1 1 1 1 0 0 1;
0 0 1 0 0 0 0;
0 0 0 0 0 0 0 ];
% Eigenvalues and eigenvectors of transpose A'
[Q1,D1] = eig(A1')
% Eigenvalues of A'
real(diag(D1))
% Principal eigenvector of A'
real(Q1(1:7,2))
% SC as diagonal of matrix exponential
SC1 = diag(expm(A1))
% Total subgraph communicability of transpose A'
sum(expm(A1')')'
% Estrada index of G, EE(G)
EE = trace(expm(A))
% Natural connectivity of A
log(trace(expm(A))/6)
% Perturbation of G by (inverse) mask M
M = [
0 0 0 0 1 0;
0 0 0 0 1 0;
0 0 0 0 1 0;
0 0 0 0 1 0;
0 0 0 0 1 0;
0 0 0 0 1 0 ];
% Natural connectivity of perturbed G
log(trace(expm(A .* (1-M)))/6)
```

REFERENCES

[1] William Ross. Ashby, *Design for a brain*, Chapman & Hall, 1952.

[2] Michele Benzi and Christine Klymko, 'Total communicability as a centrality measure', *Journal of Complex Networks*, **1**(2), 124–149, (2013).

[3] José Antonio de la Peña, Ivan Gutman, and Juan Rada, 'Estimating the estrada index', *Linear Algebra and its Applications*, **427**(1), 70 – 76, (2007).

[4] E. Estrada, 'Characterization of 3D molecular structure', *Chemical Physics Letters*, **319**, 713–718, (March 2000).

[5] E. Estrada and J. A. Rodríguez-Velázquez, 'Subgraph centrality in complex networks', *Physical Review E*, **71**(5), (May 2005).

[6] Ernesto Estrada and Desmond J. Higham. Network properties revealed through matrix functions, 2008.

[7] Ernesto Estrada, Desmond J. Higham, and Naomichi Hatano, 'Communicability betweenness in complex networks', *Physica A: Statistical Mechanics and its Applications*, **388**(5), 764 – 774, (2009).

[8] W. Grey Walter, *The Living Brain*, Gerald Duckworth & Co. Ltd., 1953.

[9] W. Grey Walter. Machina speculatrix notes on operation. http://davidbuckley.net/DB/HistoryMakers/GreyWalter/HowTortoiseWorks/, 1960. [Online; accessed 23-January-2018].

[10] M. Grinfeld, *Mathematical Tools for Physicists*, Encyclopedia of Applied Physics, Wiley, 2015.

[11] Reuben Hoggett. W. Grey Walter and the Festival of Britain (1951). http://cyberneticzoo.com/cyberneticanimals/w-grey-walter-and-the-festival-of-britain-1951/, 2015. [Online; accessed 23-January-2018].

[12] Wu Jun, Mauricio Barahona, Tan Yue-Jin, and Deng Hong-Zhong, 'Natural connectivity of complex networks', *Chinese Physics Letters*, **27**(7), 078902, (2010).

[13] H.R. Maturana and F.J. Varela, *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel Publishing Company, 1980.

[14] Victor Rosenberg, 'Opinion paper. the scientific premises of information science.', *JASIS*, **25**(4), 263–269, (1974).

[15] Yilun Shang, 'Perturbation results for the estrada index in weighted networks', *Journal of Physics A: Mathematical and Theoretical*, **44**(7), 075003, (2011).

[16] F.J. Varela, *Principles of Biological Autonomy*, Developments in Marine Biology, North Holland, 1979.

[17] J. Wu, M. Barahona, Y. Tan, and H. Deng, 'Robustness of Regular Graphs Based on Natural Connectivity', *ArXiv e-prints*, (December 2009).

Building machines that learn and think about morality

Christopher Burr [1] and **Geoff Keeling** [2]

Abstract. Lake et al. [30] propose three criteria which, they argue, will bring artificial intelligence (AI) systems closer to human cognitive abilities. In this paper, we explore the application of these criteria to a particular domain of human cognition: our capacity for moral reasoning. In doing so, we explore a set of considerations relevant to the development of AI moral decision-making. Our main focus is on the relation between dual-process accounts of moral reasoning and model-free/model-based forms of machine learning. We also discuss how work in embodied and situated cognition could provide a valuable perspective on future research.

1 Introduction

Following recent theoretical developments in deep learning, researchers have started to consider how these technologies could be leveraged to help us understand the workings of the human mind (e.g. [51]). In a recent *Behavioural and Brain Sciences* article [30], however, Lake et al. argue that "despite rapid progress in AI technologies over the last few years, machine systems are still not close to achieving human-like learning and thought." Furthermore, they state that scaling-up current systems, or utilising more data, will not be sufficient to achieve human-like learning in AI, because fundamental ingredients of human cognition are currently missing. In this paper, we explore the proposal put forward by Lake et al.[3], focusing on a specific component of mind and intelligence they do not consider directly: our capacity for moral reasoning.

Our capacity for moral reasoning is influenced by the technologies we use. It will be further shaped by ongoing technological developments, such as those discussed by Lake et al. Therefore, building machines that "learn and think" like humans, as Lake et al. propose, raises important questions about the nature of morality and our capacity for moral reasoning. For example, how will our moral decision-making and deliberation be impacted when conducted alongside artificial moral agents? And, what directions should we pursue and avoid when designing artificial agents that learn and think (like humans) about morality? We discuss some of these questions, and look at possible research directions that we believe should be critically discussed by both philosophers and those directly engaged with the research and design of AI technologies.

Despite being speculative in nature, our article avoids considering extreme possible future scenarios (e.g. superintelligence), such as those made famous by the works of philosophers such as Nick Bostrom [3]. We believe many of the most important ethical challenges surrounding AI and morality are already upon us, and require careful interdisciplinary cooperation if we are to avoid the foreseeable quagmires inherent in our current and future moral landscapes. This is important to note, for it is far easier to construct a thought experiment concerning a distant future moral scenario than it is to plan for the way that AI will actually evolve. We believe there is good reason to focus on the present and the immediate future, and to take seriously proposals such as the one defended by Lake et al., as it is research such as this that gives us the clearest indication of what the future may have in store for us.[4]

The paper proceeds as follows. In section 2, we introduce the main claims defended by Lake et al., specifically their emphasis on the use of causal models in generalisable learning, and the distinction between model-free and model-based methods of learning. We briefly mention how these topics are to be connected to the discussion of moral reasoning. In section 3, we briefly discuss what aspect of moral reasoning we focus on, and give examples from the area of moral psychology to illustrate. We also connect this section's discussion to the distinction drawn in section 2, and explain why it is relevant to the prospect of building machines that learn and think about morality. In section 4, we introduce a more philosophical consideration about the representational requirements of internal model-building, and ask whether the proposal defended by Lake et al. could be further developed by considering work in situated and embodied cognition. We conclude, in section 5, by outlining a number of ethical issues that arise at the intersection of artificial intelligence and morality.

2 Building machines that learn and think like people

In their Behavioural and Brain Sciences target article [30], Lake et al. outline a research strategy, which they believe will help engineers to develop machines that "learn and think like humans". Their strategy focuses on three non-exhaustive[5], but core ingredients of human intelligence:

1. An ability to learn and build *causal models* of the world to support explanation and understanding, rather than merely solving pattern recognition problems.
2. Grounding this learning in *intuitive theories of physics and psychology* to support and enrich the knowledge that is acquired.
3. Harnessing compositionality and learning-to-learn to rapidly acquire and *generalize knowledge to new tasks and situations*.

[1] University of Bristol, Department of Computer Science, email: chris.burr@bristol.ac.uk

[2] University of Bristol, Department of Philosophy, email: gk16226@bristol.ac.uk

[3] Although we focus on the version of their proposal defended in [30], this account is an extension of the author's earlier work, known as 'Bayesian program learning' (see [29, 45])

[4] Nevertheless, we believe there is significant value in work such as Bostrom's. We simply choose not to adopt this strategy ourselves.

[5] In their author's response, Lake et al. acknowledge that many other faculties may also be required to enable machines to fully think and learn like humans, including emotions, embodiment and action, social learning and interaction, open-ended learning, and intrinsic motivation.

To demonstrate the importance of these ingredients they discuss two recent examples of state-of-the-art deep learning systems (see [29, 33]), which are trained on two separate tasks (i.e. handwritten character recognition and generation, and learning to play video games), but drastically differ from humans in terms of key performance indicators (e.g. poor transfer of domain-general knowledge; long training periods and large datasets). Lake et al. argue that current dominant approaches in machine learning are too entrenched in pattern recognition approaches, and fail to harness more human-like methods of learning, in order to transfer acquired knowledge to new domains. For example, they state:

> "A deep learning system trained on many video games may not, by itself, be enough to learn new games as quickly as people. Yet, if such a system aims to learn compositionally structured causal models of each game—built on a foundation of intuitive physics and psychology—it could transfer knowledge more efficiently and thereby learn new games much more quickly." (p. 18)

This idea reflects an assumption made by the authors that "the difference between pattern recognition and model building [...] is central to our view of human intelligence". As an example, they consider a video game called 'Frostbite'. This video game is notoriously hard for a typical deep learning (pattern recognition) system to learn [33], due to the need for long-term planning and complex hierarchically-structured goals (e.g. acquiring a series of items before a reward is offered). Even more recent versions of deep Q-networks, which eventually outperform a human player, require hundreds of in-game hours to achieve such performance. In contrast, most human players can achieve reasonable levels of performance in a matter of minutes. Furthermore, Lake et al. state that, once learned, a human player would be able to transfer prior knowledge about the game's causal structure to *novel scenarios* (e.g. novel game mechanics such as "Get the lowest possible score", or "Die as quickly as you can") very quickly. Importantly, these novel scenarios would represent drastic departures from the initial rewards learned from prior experience, and thus represent difficult hurdles for many deep learning systems designed through standard reinforcement learning techniques. Lake et al. argue that the knowledge transfer humans display likely relies on the existence of a constructed internal model, which represents a generalisable causal structure about the game's mechanics, and is leveraged by inductive biases inherent in human learning (so called "start-up software")[6].

In spite of the strong emphasis on model-based learning, Lake et al. also discuss model-free methods of learning. In reinforcement learning, a model of the environment is an optional element in an agent's *control policy*, where the policy is alone sufficient for determining behaviour [44]. Therefore, some artificial agents can act on the basis of model-free algorithms that directly learn a control policy without needing to build a model of their environment (i.e. reward and state transition distributions). However, such agents would require a model in order to undertake more complex forms of reasoning, such as long-term planning. As is well known in artificial intelligence, building a model of the environment can be costly and time-consuming, but as the above example highlights, model-free methods are inflexible outside of highly controlled domains, making them poor candidates for generalisable learning and knowledge transfer. Therefore, as Lake et al. argue, an agent that could make use of either cooperative or competitive mechanisms for switching between model-free and model-based forms of learning (see [12]), would appear to have an advantage over less flexible agents. Such an agent would also be closer to achieving more human-like forms of learning, as existing research suggests humans are capable of utilising both model-based and model-free methods of learning (e.g. [17, 38]).

This flexible switching between model-free and model-based forms of reasoning and learning is important for understanding how the above proposal connects to moral reasoning. In section 3, we will explore the application of dual-process theories of judgement and decision-making (e.g. [27]) to accounts of moral reasoning (e.g. [19, 47]). These theories claim that in addition to relying on deliberative, model-based forms of reasoning, human agents also rely on model-free heuristics that allow the agent to trade-off accuracy for speed, while potentially selecting value-enhancing actions in constrained environments [16]. Dual-process theories are ubiquitous in the sciences of human decision-making (e.g. behavioural economics), and are also common in evolutionary psychology where they are deployed as possible adaptationist explanations for a wide-range of observed behaviours in humans and primates [49, 14]. Prior to this discussion, however, it is important to address a theoretical assumption.

The perspective that Lake et al. adopt is explicitly computational in nature—that is, intelligent behaviour can be causally explained by appealing to a series of algorithmic processes that the agent's cognitive system realises [37].[7] However, Lake et al. also acknowledge, that this is unlikely to be sufficient to capture all forms of human intelligence:

> "Other human cognitive abilities remain difficult to understand computationally, including creativity, common sense, and general-purpose reasoning." (p.3)

In section 4, we will discuss a possible research avenue, inspired by work in social cognition and embodied robotics, which argues for the importance of cognitive scaffolds and niche-construction in supporting adaptive behaviour. We will argue that one possible hurdle that computational approaches could face, may arise with an implicit commitment to a methodological individualism, which views the brain's mechanisms as the primary system to be explained (in computational terms) when we wish to understand an agent's behaviour (see [9] for discussion). It is unclear to what extent Lake et al. are committed to a methodological individualism[8], and so our proposal is intended as a friendly suggestion that we believe is in the spirit of their account.

In contrast to the kind of methodological individualism that characterised classical cognitivist approaches to mind and behaviour, recent work in '4e cognition'[9] argues that some forms of human (and non-human) intelligence arise from an agent's engagement with its material environment (e.g. [32]) and embeddedness within its socio-

[6] Lake et al. acknowledge that human learning has itself been shaped by millions of years of evolution, which could be seen as our own "training period". However, this point merely reinforces their argument for developing a similar type of "start-up software" for artificial agents, which natural selection has developed for humans.

[7] Although we believe it is also worth considering whether moral reasoning could be better explained in non-computational terms, we restrict ourselves in this paper to considering research that is most directly relevant to the computational approach being discussed.

[8] For example, in their author's response, Lake et al. state that their intention was to remain agnostic about possible implementations for how models should be learned (see section R5.2 in [30]).

[9] '4e cognition' refers to work that fits within the research programmes of embodied, embedded, extended, and enactive cognition. It does not represent a unified research programme itself.

cultural niche (e.g. [2]).[10] The engagement between an agent and its environment can include the leveraging of physical structures (including the agent's own body) to reduce the computational demand that a given task places on the agent's cognitive system (e.g. reordering ingredients in a recipe, in order to reduce the demands on an agent's memory), but can also extend to normative constraints that are embedded within social institutions (e.g. language, legal structures, social norms). These normative constraints may themselves provide readily accessible alternatives to the costly construction of an internal model (e.g. using emotional feedback from peers as an approximate indication of whether your actions are socially acceptable).

In the following sections, we expand on each of the above points, which we believe offer fruitful ways of thinking about how to build artificial systems that could be capable of rudimentary forms of moral reasoning, or perhaps better support existing forms of human moral reasoning (e.g. "human-in-the-loop"). However, before we discuss these features, it is worthwhile stating what we mean by 'moral reasoning'.

3 What Is 'Moral Reasoning'?

When we ask what would be required for an AI to think and learn about morality, we must be clear about the kind of moral reasoning in question. There are, at least, three kinds of cognitive process which might reasonably be classed as 'moral reasoning'. In this section, we distinguish these different kind of moral reasoning, and make clear which of these is under consideration.

First, moral reasoning might be understood as the kind of reasoning demanded by the correct normative ethical theory (e.g. if utilitarianism is the correct theory, then moral reasoning is reasoning in accordance with utilitarianism). Second, moral reasoning might be characterised as a form of deliberation which requires us to adopt an *impartial perspective*. That is, moral reasoning requires us to consider the interests of a suitable reference class of moral patients, as opposed to just our own interests [41, 23]. Finally, according to a third, so-called *descriptive view*, moral reasoning might be understood as reasoning which involves *moral concepts*, such as fairness, duty, blame and responsibility. The focus of this third view is how humans do reason about morality, as opposed to how they ought to reason. As Lake at el. are primarily concerned with what is required to bring AI closer to human cognition, we focus on this third view.

The *descriptive view* involves a commitment to two claims. The first is that our moral concepts are built around innate tendencies to evaluate certain features of our environment [43, 26]. These evaluative tendencies admit evolutionary explanations. For example, take the disposition to evaluate characteristically *unfair* situations as bad. Plausibly, natural selection favoured genes promoting emotional dispositions *against* unfair situations, as these dispositions serve an important regulatory function that allows organisms to reap the benefits of prosocial behaviour. It is, therefore, no surprise that other primates have negative emotional responses to characteristically unfair situations [5, 6]. The second commitment involves the role of folk-psychological concepts in our moral reasoning. Guglielmo, Monroe and Malle [20], for example, have argued that many of our most important moral concepts are grounded in folk-psychology. Our concept of blame, for example, relies on our seeing the recipients of blame as *agents* capable of intentional action, foreseeing consequences, and so on. Joshua Knobe [28] defends a related thesis, according to which our moral concepts are central to how we understand intentional action. This commitment connects up with the first of the two insights from Lake et al., and we take it to be a positive feature of their argument that they acknowledge the relevance of grounding causal models in intuitive theories of folk-psychology in human cognition, insofar as this may help to capture important features of our moral reasoning. However, the types of models that Lake et al. emphasise are richly-structured generative models, which work by trying to reconstruct the hidden causal structure of a target domain (e.g. perception), and it is unclear to what extent this theory-like model-building is a necessary of our capacity for moral reasoning.

As alluded to in section 2, Lake et al. deal with this worry by reference to a key debate in reinforcement learning: the extent to which an intelligent agent relies on model-free or model-based methods of learning and decision-making. They acknowledge that some task domains are best approached using model-based methods of cognition (e.g. deliberative planning), whereas others seem to require model-free methods (e.g. skillful or habitual motor activity), and that some recent proposals in artificial intelligence and computational neuroscience use a combination of the two (e.g. [38, 48, 34]). The extent to which a particular task requires model-based or model-free methods of cognition is likely a matter of degree, and may require some sort of arbitration mechanism to alter the extent to which the two forms interact (see [10]). Regardless, neuroscientific evidence supports the idea that human learning comprises both model-free and model-based methods [17, 11]. How does this matter for moral reasoning?

If our moral concepts are grounded in our folk psychology, as Guglielmo, Monroe and Malle [20] argue, then one way of understanding the model-based versus model-free distinction, is as a guide to when our moral reasoning relies *most heavily* on deliberative or habitual processes[11]. Joshua Greene (e.g. [19, 18]) argues that moral reasoning involves an interplay between affective or 'quick-fire' cognitive processes and our deliberative cognitive processes, and this aspect of our morality may be nicely captured by the second insight from Lake et al. (i.e. an interplay between model-free and model-based learning). In his [18], Greene found empirical evidence showing that the way in which a moral dilemma is presented to us influences our deliberation about that dilemma. For example, in trolley cases, we are inclined to kill the one to save five if doing so involves causing the harm 'remotely' (e.g. pulling a leaver). But in cases which involve 'up close and personal harm' we are liable to have a negative emotional response which biases our deliberations in favour of letting the five die so that we avoid inflicting harm on the one.

This interplay can help us overcome some of the worst effects of using heuristic (or model-free) based forms of reasoning. As is well known, heuristics are often adaptive only in narrow domains [16], and there is some reason to think that heuristics make us *worse* moral reasoners outside of these constraints. Greene [18] and Peter Singer [40] have argued that the role of heuristics in moral reasoning causes us to be sensitive to morally irrelevant features of decision problems. For example, we are moved to help individuals suffering nearby to

[10] In some cases, the account that is offered rejects a computational perspective, in favour of a more dynamical approach (e.g. [8]).

[11] As already alluded to, it is likely that the extent to which certain forms of reasoning and learning are best described as "model-free" or "model-based" is a matter of degree. Therefore, it is ill advised to assume that the traditional dichotomies between habit and reason, or heuristics and deliberation, map neatly onto the distinction between model-free and model-based.

us, but not on the other side of the world, yet, according to Singer [40], the location of an individual is irrelevant to whether we ought to help them. In light of this, the proposal of Lake et al. to develop artificial systems that are able to adaptively deploy both model-free and model-based forms of learning and reasoning appears sensible in light of this worry.

In this section, we distinguished three different accounts of *moral reasoning* and specified the account which we intend to focus on. The descriptive accounts of moral cognition found in moral psychology will be the object of our inquiry. In what follows, we explore how model-free and model-based forms of learning, alongside embodied and situated cognition, can elucidate what it would mean for an AI to think and learn about morality.

4 The world as its "own best model"

In his [4], Rodney Brooks, offered a criticism of what he termed the 'sense-model-plan-act' (SMPA) model of artificial intelligence. The idea that Brooks wished to challenge was that if an AI (or a robot) was a) required to gather information from its environment (sensing), in order to b) build a richly reconstructive representation (model), with which to c) formulate a plan of reaching some desired goal-state (plan), before d) carrying out the necessary movements (action), then outside of a carefully designed and controlled laboratory setting (i.e. a narrow domain), such a serial process would be insufficiently dynamic to cope with the pressures of a constantly changing environment. In the time taken to deliberate, the environment may have changed, rendering the current model (and any actions based on it) inaccurate, and thus raising the agent's uncertainty. Utilising the SMPA model in ecologically-valid scenarios would mean either the artificial agent would incur an accuracy cost (subject to the environment changing), or it would incur a drastic speed cost. Instead, Brooks' suggestion was to implement a more straightforward sensorimotor coupling approach (based on his *subsumption architecture*), where the internal models were replaced with a more direct sensitivity to the environment, and the environment directly elicited and constrained adaptive behaviour with no need for mediating representations.

Since this time, greater consideration has been paid to the speed-accuracy trade-off, and the distinction between model-free/model-based methods has evolved to a point where many researchers now acknowledge the importance of some type of arbitration mechanism between the two methods (e.g. [10, 12]), rather than accepting a strict dichotomy. However, as some of the commentators to the Lake et al. target article argued, more attention still needs to be given to more ecologically-valid forms of intelligence that rely on the agent's situatedness or embodiment (e.g. [1, 35]). In short, if the body or environment of the agent enables the agent to offload some of the computational complexity, then there may be no need for the agent to construct a detailed inner model of the environment in the first place—in Brook's own words, "The world is its own best model." (1991, p. 15). This idea is reflected in work in developmental psychology [46] and soft robotics [36], and, in some cases, represents an instance of what Robert Wilson [50] refers to as 'wide computationalism'. It is also a familiar research area discussed in the 'extended mind' literature. In this section, we extend some of these considerations to the issue of moral reasoning—an area that is often underexplored in the 4e cognition literature.

One way of understanding the embodied and situated cognition research programmes, when applied to moral reasoning, is in uncovering the myriad ways that our environment (including our bodies) shapes and constrains the way we learn and reason about the world. Our environment represents an irreducible source of uncertainty and complex hierarchically-structured causes (e.g. what consequences will my actions have on other agents worthy of moral consideration), and our brains have clearly evolved heuristics and biases in order to simplify some of this complexity [16]. In acknowledging this, embodied and situated cognition researchers point to the way that social interaction allows us to cooperatively shape our sociocultural niche (e.g. [42]), and possibly make the world more predictable by constructing a more reliable domain in which our heuristics can operate (i.e. intervening on the world to reduce uncertainty). More recently, researchers in the area of *normative folk psychology*, have presented evidence for how certain sociocultural norms (including morality) are constructed through social interactions, and in turn contribute to our understanding of our own behaviour [52]. The benefit of constructing a stable, normatively structured environment is not only that it helps to regulate behaviour, but also that it provides a way of offloading some of the computational demands of cognition onto the environment itself. Acknowledging when this is possible (and desirable) could help AI researchers determine when artificial systems need to rely on model-based methods, or when the world can stand-in as "it's own best model". In the case of morality, by paying attention to the structure of the environment, engineers can determine if some normative structure is already present, and whether it is better to simply couple an agent's actions to the world as a sort of *distributed form of moral behaviour*. To better make sense of this, consider the following example.

H. L. A. Hart [24] provides an account of what distinguishes societies with a legal system from societies without a legal system. According to Hart, a set of norms becomes a legal system when 'secondary legislation' is enacted which stipulates the conditions under which a rule ought to be recognised as law. For example, the constitution of a state will specify which individuals are permitted to enact valid laws. If Hart's view is correct, then the development of secondary legislation could plausibly be construed as an instance of cognitive offloading with respect to moral cognition. Secondary legislation provides individuals in a society with a *prima facie* reason to behave in accordance with primary legislation, even if they do not understand the argument behind the primary legislation. Put another way, once secondary legislation is introduced, individuals have reason to comply with certain imperatives *because it is the law*. Thus, the existence of secondary legislation provides an external constraint on our legal (and often moral) behaviour, which does not require us to evaluate whether or not there is good reason to comply with the constraint.

The above example should highlight that a moral agent is not required to *internalise the norms* of society in order to ensure their behaviour meets certain moral standards, and can potentially make do with a simplified model of the world (or maybe even a set of well-tuned heuristics) when certain institutions act as regulative constraints. Of course, delineating the causal factors that govern an agent's behaviour is understandably a complex task. However, it is important to realise when an agent may be able to behave optimally (e.g. morally) simply by utilising adaptive heuristics, which respond to simple cues in the environment, rather than by constructing a rich inner model that acts as the basis for deliberative decision-making. This is not to deny that human agents are capable of norm internalisation, but the extent to which our moral behaviour is a product of constrained heuristics, rather than model-based deliberation is unclear. For example, it is possible that we achieve a high degree of moral optimality by using model-based reasoning (perhaps implemented by mechanisms in prefrontal cortex) to competitively constrain the more heuristic based forms of action selection that drive our moral behaviour. However, it is also likely that society has collectively shaped our shared sociocultural niche, in order to reduce the demands placed on individuals, while nevertheless promoting optimal decisions.

An important question remains, why should we design artificial agents that rely on (potentially maladaptive) heuristic decision-making, like humans, when there is the possibility of pursuing more rational methods. Is there an answer, short of avoiding computationally demanding model-building, that can be offered?

5 Further Remarks

We conclude, by briefly considering whether the development of AI systems that are capable of human-like moral reasoning is a desirable goal. Our aim is not to argue exhaustively in favour of either positions, but rather to provide a sketch of the related ethical challenges that arise at the intersection of artificial intelligence and morality. We begin with the negatives.

5.1 Why artificial morality may be undesirable

(1) *The Ideal Reasoner Concern*: It might be the case that AIs which reason morally as *we* do are more inclined to make suboptimal moral decisions. As aforementioned, heuristics sometimes make us sensitive to morally irrelevant features of decision problems. So, if our intention is to develop AIs which make the *best possible* moral decisions, then we might have stronger reasons to focus on building *ideal* moral reasoners, as opposed to AIs which replicate our non-ideal moral reasoning. But, in order to design an ideal moral agent, we need to have a clear picture of what an ideal moral agent looks like [25, 3]. There are at least two problems here. On the one hand, there are several plausible ethical theories on the market, and moral philosophers are yet to provide decisive reason to favour one of these theories. Moral philosophers have only recently started to consider the rational response to 'normative uncertainty', providing decision-theoretic accounts of how to adjudicate between moral theories when we are unsure which, if any, is correct [31]. So, it is unclear at present which moral principles we have best reason to implement when designing an ideal moral agent. On the other hand, it is often the case that apparently plausible moral principles give surprising results in novel situations. Indeed, a substantial amount of ethical theorising involves testing how different principles square with our intuitions in novel cases. Whilst a set of moral principles might seem 'ideal' in one setting, they can easily be non-ideal in another. So, in developing AIs as 'ideal' moral reasoners, it is plausible that the principles used might deliver unforeseen and counterintuitive results, which is something we have good reason to be cautious about [3].

(2) *The Bias Concern*: Jonathan Haidt [22] notes the importance of in-group/out-group biases in our moral decision making.[12] Plausibly, whilst this bias may have an important functional role in human moral reasoning, there is good reason not to include biases of this kind in our AIs. The problem is that there exists a gap between how humans *in fact* reason morally given the available cognitive mechanisms, and how humans *ought to* reason morally. This gives rise to a trade-off. On the one hand, designing AIs to reason about morality as humans do will make AIs susceptible to the kinds of moral mistakes that humans routinely make. Human moral reasoning is imperfect, at least insofar as our cognitive mechanisms have inbuilt biases which dispose us to factor in morally irrelevant information into our decision-making. On the other hand, designing AIs to reason morally in a way that is too far removed from ordinary human reasoning will most likely result in AIs with inflexible moral reasoning that is unsuitable for general use across a broad class of moral decision-problems.

There are some biases in human cognition which would provide no straightforward benefit to AI moral reasoning if analogous biases were implemented into AIs. Furthermore, the non-inclusion of these biases is unlikely to be problematic. Consider *ego depletion*. According to what is called the *strength model*, humans have a capacity for self-control which enables them to engage in goal-directed behaviour and to resist impulses. However, this capacity is limited: prolonged activity involving self-control diminishes our capacity to resist impulses [21]. Although ego depletion no doubt plays a role in human moral decision-making, which is a taxing exercise requiring considerable self-control, there are good reasons not to include an analogous bias when developing AIs to reason morally. There is no clear benefit to having AIs which are in some way *worse* at making moral decisions after a prolonged period of moral decision-making. And it is desirable that AI moral reasoning is *consistent* over time. (This is especially important with respect to AIs making decisions which affect human wellbeing, such as AIs used to aid decision-making in criminal justice.) In our view, biases like ego depletion ought not to be considered as *necessary* for constructing an AI which reasons about morality, even though human moral reasoning is no doubt afflicted by this cognitive limitation.

We have examined two concerns about the desirability of AIs which have the capacity for moral reasoning. In the next section, we discuss two desirable features of AIs capable of moral reasoning.

5.2 Why artificial morality may be desirable

(3) *The Transparency Concern*: With regards to positive reasons for pursuing artificial morality, we first consider the issue of transparency in current deep learning systems [7]. The transparency concern relates to artificial decision-making systems that process information using methods that are opaque to most people affected by the AI's decision. It is, therefore, difficult to provide an *explanation* of how the AI reached its decision. Explanations are an important component of how we engage with other moral agents in society. Indeed, some moral theorists, such as T.M. Scanlon [39], argue that what makes actions morally right or wrong is whether those actions are mandated by principles which are justifiable to the parties affected by those principles. On this view, reasons take centre stage, as we justify our moral behaviour by appealing to reasons. In principle, the issue of transparency could be resolved if we develop AIs whose moral reasoning is grounded in folk-psychological mechanisms similar to our own moral reasoning. As Jonathan Haidt [22] and others have noted, when we attempt to justify our moral behaviour, these

[12] Note that there are three kinds of bias that present ethical issues in the context of AI decision-making: (1) We sometimes claim that a dataset is biased. When bias is a property of datasets, we mean that the sample data does not accurately represent the population. This kind of bias can present ethical issues with respect to, at least, training data for Artificial Neural Networks (ANNs). For example, an ANN used for voice recognition might be trained on a dataset of voices in which minority accents are insufficiently represented. (2) Other times we claim that a decision-making process is biased. For example, an ANN which is trained on biased data might produce skewed classifications. (3) Yet more times, we say that a decision-maker has a bias which is part of the agents cognitive apparatus. For example, an agent could have in-group/out-group bias as a component of their cognitive apparatus.

justifications involve folk-psychological concepts such as intention, belief and reasonable foresight of consequences. In bringing AIs in line with our mechanisms for moral reasoning, plausibly, this will open the possibility of AIs who can *themselves* offer moral justifications for decisions which are intelligible to those affected by the AI's decision. Importantly, as Daniel Dennett [13] notes in the case of the *Intentional Stance*, these types of explanations need make no reference to the underlying mechanisms that ground an agent's behaviour (e.g. the pattern of neural activity that causes an agent's actions), which is important given the inherent opacity of (black-box) deep learning systems. Explanations that are couched in terms of intentional psychological states (e.g. beliefs, desires etc.) play a simplifying role, which can also have a regulative effect on our future behaviour, and are typically sufficient to justify moral behaviour. For example, we are presumably happy for someone to justify their behaviour by virtue of appeal to folk psychological states, rather than a more complex explanation that makes reference to neural states[13]. Building artificial agents whose learning is grounded in intuitive folk psychological theories, as Lake et al. propose, seems a sensible first step in working towards artificial intelligence more understandable to humans.

(4) *The Envelope Concern*: Secondly, and finally, building artificial systems that can (co)operate within our own system of moral values is important, as we ideally want to avoid developing intelligent systems that are misaligned with our own moral principles. Research in situated and embodied cognition may represent a valuable avenue to explore in this regard, allowing us to develop autonomous decision-making systems that cooperate with us and help us overcome some of the limitations of our own moral reasoning. Luciano Floridi's [15] notion of an 'envelope' is a helpful conceptual tool to understand this point. He states, "In robotics, an *envelope* is the three-dimensional space that defines the boundaries that a robot can reach. We have been enveloping the world for decades without fully realising it." Here, the problem is that by "enveloping our world" such that it is easier for artificial agents to operate within it, we end up restructuring our own environment in ways that may have problematic consequences for us. We do not want our environment (including ourselves) re-structured to fit the ontology and values of artificial agents that may have conflicting goals or morals. In short, we want to design artificial agents that can (co)operate within our own envelope, not change our environment to fit theirs. Of course, the process will likely be a matter of reciprocal development (e.g. encoding knowledge systems in a AI-friendly manner; re-designing roads to accommodate autonomous vehicles), and humans are able to adapt to new situations thanks to our ability to learn generalisable knowledge. Pursuing artificial agents that "learn and think" more like us, however, may help make the process more conducive to human flourishing.

ACKNOWLEDGEMENTS

Christopher Burr is supported by a European Research Council Project [ThinkBIG, Advanced Grant (AdG), PE6, ERC-2013-ADG], awarded to Nello Cristianini. Geoff Keeling's contribution to this work was funded by an Arts and Humanities Research Council PhD Studentship, awarded through the South, West and Wales Doctoral Training Partnership.

[13] Of course, there may be exceptions to this. For example, a legal case in which a defendant appeals to an underlying defect in the neurophysiology, as an excuse for their behaviour. In these instances, folk psychological explanations may only be one component of the defendant's full justification.

REFERENCES

[1] Gianluca Baldassarre, Vieri Giuliano Santucci, Emilio Cartoni, and Daniele Caligiore, 'The architecture challenge: Future artificial-intelligence systems will require sophisticated architectures, and knowledge of the brain might guide their construction', *Behavioral and Brain Sciences*, **40**, (2017).

[2] Louise Barrett, *Beyond the brain: How body and environment shape animal and human minds*, Princeton University Press, 2011.

[3] Nick Bostrom, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, 2016.

[4] Rodney A Brooks, 'Intelligence without representation', *Artificial intelligence*, **47**(1-3), 139–159, (1991).

[5] Sarah F Brosnan and Frans BM De Waal, 'Monkeys reject unequal pay', *Nature*, **425**(6955), 297, (2003).

[6] Sarah F Brosnan and Frans BM de Waal, 'Evolution of responses to (un) fairness', *Science*, **346**(6207), 1251776, (2014).

[7] Jenna Burrell, 'How the machine thinks: Understanding opacity in machine learning algorithms', *Big Data & Society*, **3**(1), 2053951715622512, (2016).

[8] Anthony Chemero, *Radical Embodied Cognitive Science*, MIT press, 2011.

[9] Anthony Chemero and Michael Silberstein, 'After the philosophy of mind: Replacing scholasticism with science', *Philosophy of science*, **75**(1), 1–27, (2008).

[10] Andy Clark, 'The many faces of precision (replies to commentaries on whatever next? neural prediction, situated agents, and the future of cognitive science)', *Frontiers in psychology*, **4**, 270, (2013).

[11] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan, 'Model-based influences on humans' choices and striatal prediction errors', *Neuron*, **69**(6), 1204–1215, (2011).

[12] Nathaniel D Daw, Yael Niv, and Peter Dayan, 'Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control', *Nature neuroscience*, **8**(12), 1704, (2005).

[13] Daniel Clement Dennett, *The intentional stance*, MIT press, 1989.

[14] Jonathan St BT Evans, 'Dual-processing accounts of reasoning, judgment, and social cognition', *Annu. Rev. Psychol.*, **59**, 255–278, (2008).

[15] Luciano Floridi, 'Enveloping the world for ai', *The Philosophers' Magazine*, (54), 20–21, (2011).

[16] Gerd Gigerenzer and Reinhard Selten, *Bounded rationality: The adaptive toolbox*, MIT press, 2002.

[17] Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty, 'States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning', *Neuron*, **66**(4), 585–595, (2010).

[18] Joshua D Greene, 'The secret joke of kants soul', *Moral psychology*, **3**, 35–79, (2008).

[19] Joshua D Greene, R Brian Sommerville, Leigh E Nystrom, John M Darley, and Jonathan D Cohen, 'An fmri investigation of emotional engagement in moral judgment', *Science*, **293**(5537), 2105–2108, (2001).

[20] Steve Guglielmo, Andrew E Monroe, and Bertram F Malle, 'At the heart of morality lies folk psychology', *Inquiry*, **52**(5), 449–466, (2009).

[21] Martin S Hagger, Chantelle Wood, Chris Stiff, and Nikos LD Chatzisarantis, 'Ego depletion and the strength model of self-control: a meta-analysis.', *Psychological bulletin*, **136**(4), 495, (2010).

[22] Jonathan Haidt, 'The new synthesis in moral psychology', *science*, **316**(5827), 998–1002, (2007).

[23] John C Harsanyi, 'Morality and the theory of rational behavior', *Social research*, 623–656, (1977).

[24] Herbert L A Hart, 'The concept of law', (1961).

[25] Michael Hauskeller, *Better humans?: Understanding the enhancement project*, Routledge, 2014.

[26] Richard Joyce, *The evolution of morality*, MIT press, 2007.

[27] Daniel Kahneman, *Thinking, Fast and Slow*, Macmillan, 2011.

[28] Joshua Knobe, 'Intentional action in folk psychology: An experimental investigation', *Philosophical psychology*, **16**(2), 309–324, (2003).

[29] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum, 'Human-level concept learning through probabilistic program induction', *Science*, **350**(6266), 1332–1338, (2015).

[30] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman, 'Building machines that learn and think like people', *Behavioral and Brain Sciences*, **40**, (2017).

[31] William MacAskill, 'Normative uncertainty as a voting problem', *Mind*, **125**(500), 967–1004, (2016).

[32] Lambros Malafouris, *How things shape the mind*, MIT Press, 2013.

[33] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., 'Human-level control through deep reinforcement learning', *Nature*, **518**(7540), 529, (2015).

[34] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine, 'Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning', *arXiv preprint arXiv:1708.02596*, (2017).

[35] Pierre-Yves Oudeyer, 'Autonomous development and learning in artificial intelligence and robotics: Scaling up deep learning to human-like learning', *Behavioral and Brain Sciences*, **40**, (2017).

[36] Rolf Pfeifer and Josh Bongard, *How the body shapes the way we think: a new view of intelligence*, MIT press, 2007.

[37] Gualtiero Piccinini, 'Computationalism in the philosophy of mind', *Philosophy Compass*, **4**(3), 515–532, (2009).

[38] Evan M Russek, Ida Momennejad, Matthew M Botvinick, Samuel J Gershman, and Nathaniel D Daw, 'Predictive representations can link model-based reinforcement learning to model-free mechanisms', *PLOS Computational Biology*, **13**(9), e1005768, (2017).

[39] Thomas Scanlon, *What we owe to each other*, Harvard University Press, 1998.

[40] Peter Singer, 'Ethics and intuitions', *The Journal of Ethics*, **9**(3-4), 331–352, (2005).

[41] Peter Singer, *Practical ethics*, Cambridge university press, 2011.

[42] Kim Sterelny, 'Thought in a hostile world: The evolution of human cognition', (2003).

[43] Sharon Street, 'A darwinian dilemma for realist theories of value', *Philosophical Studies*, **127**(1), 109–166, (2006).

[44] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, volume 1, MIT Press, 1998.

[45] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman, 'How to grow a mind: Statistics, structure, and abstraction', *science*, **331**(6022), 1279–1285, (2011).

[46] Esther Thelen and Linda Smith, *A dynamic systems approach to the development of perception and action*, Cambridge: MIT Press, 1994.

[47] Michael Tomasello, 'Why be nice? better not think about it', *Trends in cognitive sciences*, **16**(12), 580–581, (2012).

[48] Théophane Weber, Sébastien Racanière, David P Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adria Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al., 'Imagination-augmented agents for deep reinforcement learning', *arXiv preprint arXiv:1707.06203*, (2017).

[49] Andrew Whiten, 'Chimpanzee cognition and the question of mental rerepresentation', *Metarepresentation: a multidisciplinary perspective. Oxford University Press, Oxford*, 139–167, (2000).

[50] Robert A Wilson, 'Wide computationalism', *Mind*, **103**(411), 351–372, (1994).

[51] Daniel LK Yamins and James J DiCarlo, 'Using goal-driven deep learning models to understand sensory cortex', *Nature neuroscience*, **19**(3), 356, (2016).

[52] Tadeusz Wieslaw Zawidzki, *Mindshaping: A new framework for understanding human social cognition*, MIT Press, 2013.

An article conforming to the formatting of AISB 2008

Jack Robert Edmunds-Coopey

Machinics: *Philosophy of Technicity to Programming, Towards a Contemporary Ontology of Artificial Life as a new Form-of-Life and Technics, A Critique of Reactionary Modernism against Technology*

Computing systems and computational software are seen as artifacts in their own right different and similar to other forms of objects differentiated from the digital, which seems to require a philosophical analysis of its own because of its differentiated, metaphysical nature from ordinary material entities. The computer and its programming appears like a new world within a world. It is a technical object which acts like a machine which produces other forms of technical objects. However, it seems that the metaphysical boundaries and distinctions that are drawn in the philosophy of computer science are remarkably demarcated by their Enlightenment, organicist latent assumptions and presuppositions. Physical machines cause physical things to move put bluntly, but programmes cause physical things to move, to what extent is contemporary philosophy of computer science limited in its dualistic, and non-technical understanding and theorization of programming as such? The subject-object binary in the history of Western philosophy has long since been a problem and has been explored in differing ways, more specifically the tertiary relation of the object, or the technical being as intermediary and constitutive of the human subject and its objective world has been discussed from Plato to contemporary Bernard Stiegler. The combination and duality between the supposedly physical hardware and abstract software in relation to the question of programming is seemingly blended and entangled, in which a new theory of digital life is required to synthesize a contemporary understanding of these new potentialities of programming. It seems that the binaries between physical and non-physical, specification and function, semantic and syntactic implementation, abstract and concrete are too deflationary not only 'within' the machine who programs, but also in the subject who is doing the activity of programming. Thus, it is proposed in this paper that a tertiary ontology of the technical, digital form of object is required to not only undo and challenge these problematic binaries which sustain Enlightenment metaphysics between machines of stuff and physicality and programming of semantics, but also to construct a new necessitated ontology of the digital as a new, distinct mode of being separate from all other historical modes of life to come in human history. Gilbert Simondon and his philosophy of technics is suited to analysing the philosophical and anthropological consequences of programming as a type of knowledge that is not only individuated by itself, but individuates the human subject itself. Utilizing Simondon's ontology of technics using contemporary physics, one can understand the plasticity of programming as simultaneous physical, non-physical and mathematical objects in their dynamic nature as technical artifacts by themselves aside from the machines. In addition to Simondon's theory of technics, contemporary Continental philosophy of mind in relation to the work of Catherine Malabou utilizing recent discoveries in cognitive science in relation to brain plasticity and computational theories of the mind bridge onto Simondon's ontology of programming and technics by helping to decipher the relation between the mind, the brain and the machine of programming, and the relation between. In essence, the question of programming requires less now a philosophy as such, but a philosophical physics which can account for contemporary physics as well as the latest developments in programming which neither physics nor philosophy can assess because of its complex nature of material and metaphysical properties which archaic dualities such as concrete and abstract cannot answer. In essence, this paper wishes to problematize the ontology of programming as a distinct form of life unlike any other, in which the function of computational artifacts and technical objects are absolutely differentiated to that of former models of computers. Thus, the abstraction and semantics of computers in relation to programming are distinct not only semantically from ordinary writing systems and ethics of responsibility as a result, but fundamentally cannot be understood and ontologized in the same ways previous ontologies of languages, technical objects and other forms of life that have been hitherto, theorized. Therefore, in the words of Gilles Deleuze, the ontology of programming offers a 'new line of thought' where computer scientists, philosophers, historians and physicists alike can coalescence in attempting to grasp the nature and potentiality of the thingness of programming itself whether it floats between physical, non-physical, abstract, concrete, virtual or actual.

> From another side: is Achilles possible with powder and lead? Or the Iliad with the printing press, not to mention the printing machine? Do not the song and the saga and the muse necessarily come to an end with the printer's bar, hence do not the necessary conditions of epic poetry vanish?[1]

I started with this quotation from Marx's Grundrisse in order to demonstrate a basis or analysing the structure between the human mind, its technical objects and the artificial as a result. Is the human mind and creativity possible with Artificial Intelligence? With the emergence of artificial intelligence, does Marx's analysis still ring true? What is the program and its relation to the human mind in light of artificial intelligence? What I wish to propose today is a Hegelian solution to this three-fold problem. To use the structure of Hegel's Phenomenology of Spirit from beginning at self-consciousness (Malabou's work on the brain/mind), to the object (Simondon's work on technics), to Nature (Stiegler's work on the technical and human relation) to provide a sketchwork of an ontology of artificial intelligence.

I wish to argue that in the advent of Artificial Intelligence it does not just problematize the notion of machinery and its advancement, but the very interstitial relations between the human mind and its externality, in Hegel's terms called Nature. It challenges the very core essence of what we could define not only as our mind and its capacities, but the very dialectical process by which the human mind makes sense of its environment, assimilates various parts of it, and produces technicities through this interaction which then change the human mind as a result. Therefore, an ontology of Artificial Intelligence in terms of its effects on the human mind, its nature and technicity is needed before we can assess a means by which to handle AI overall as a contemporary phenomenon. In Hegel's *Phenomenology of Spirit* (1807) he analyses the developments of historical self-consciousness through a series of concentric circles that mediate one another, but where is the mechanism or machine in his analyses? In the Grundrisse (1857-1858), Marx produces his famous fragment on machines which details a differing relation to machines and its effects on the distribution and inflation of capital, and its effects on the workers themselves. In this paper, I firstly wish to take account of Hegel's writings on the machine and its possible consciousness. Secondly, I wish to examine Marx's writings in the Grundrisse to analyse the Hegelian overtones in his conceptions about the ontology of the machine and its consciousness, but also its effects on the human mind as a result.

As a consequence of these two parallel readings on the question of the machine, I wish to then propose a Hegelian construction of attempting to link the three areas of contention, the brain, the technical object itself, and the relation between them by firstly sketching Catherine Malabou's work on the brain, then outline Gilbert Simondon's work on the question of technical objects, to then mediate these two spheres of the brain and the technical object with a note on Bernard Stiegler's use of technics to generate an understanding of constitutive subjectivity between the brain and the machine-object. Therefore in essence, I wish to explore an ontology of the object and its constitution through technics with an exploration of how we can conceive of the brain in this process. Firstly, an explication of Hegel's analyses concerning the mind and its relation to mechanism and machine is needed to flesh out Marx's fragment on machines. Richard Dien Winfield's article *Hegel, Mind, and Mechanism: Why Machines have no Psyche, Consciousness, or Intelligence* (2009) whose key point is that machines have modelled on, and seen to be simulated as same

[1] Marx Karl, Introduction, Late August - Mid-September 1857 (1) PRODUCTION Independent Individuals. Eighteenth-century Ideas in (fore. Trans. by Martin Nicolaus) in Grundrisse, Foundations of the Critique of Political Economy (Rough Draft) Penguin Books, New Left Review, (London, 1973), p. 111

as mental reality or seeing the mind as a mechanism of some kind.[2]

Therefore, the first part of this paper concerning Hegel is to understand how the metaphysical and ontological relation between the mind and the outside nature and its objects can be understood. Broadly, the relation between chemicals and the causality of mechanical objects is different to the relations between mind and nature which Hegel was one of the few to recognise. In addition to this, a side note in light of recent discussion of technics in contemporary Continental philosophy, the malaise and neurosis of technics is fundamentally in my view, a reactionary modernist one, having its roots in the work of Ernst Junger, Lewis Mumford and Osward Spengler, and it is curious to me how these discussions of technics have been taken up in more recent times in more progressive circles.

> As Hegel's own account of 'Objectivity' reveals, this irreducible supervenience occurs most minimally in chemical process, or chemism (EL: §§200-203, 265-67; SL: 727- 31). Like mechanism, chemical process determines objects externally without purpose or form. Just as motion gets mechanically communicated to one object by another, objects must be brought together by some external catalyst to react chemically. Chemical relations are distinguished from mechanical interaction in that objects are chemically affected by one another not as mere bodies governed by the same laws of motion, but as distinct chemicals that are poised to break down or coalesce in function of their complementary difference. The mechanical relations of objects as movable matter can neither be violated nor impeded by chemical process precisely because chemism pertains to the relational difference of objects, to which mechanism is indifferent. Whereas objects react chemically without relinquishing their governance by laws of motion applying to matter qua matter, their chemical reaction involves something undetermined by and irreducible to mechanism.[3]

Therefore, Hegel's account of chemism in the Encyclopaedia Logic and the Science of Logic attempts to conceive of the relation between the mind and nature in relation to the interstices of their cleavage. The point of analysing Hegel's chemism is to suggest that when chemicals interact they physically meet and produce reactions and 'complementary differences', but to what extent does the mind and its supposed locus, the brain interact with an apparent outside or a representation or projection? It is at this precise juncture that the question of the mind and its outside where this paper tries to answer whilst mediated by the technical object which constitutes both the outside nature and the mind.

> Hegel points out a fundamental symptom of this inability: mechanical 'thinking' always manipulates terms with a content given and fixed. What machine intelligence orders with indifference to its content is something thereby both undisturbed by those operations and at hand independently of them. If thought determinations were condemned to have the atomistic, passive, and rigid character of such inputs, concepts could never relate to what they are not, nor develop themselves into new conceptual content. This would leave thought wholly analytic, unable to generate any determinacies of its own, and reduced to an impoverished instrument for sorting what lies within given terms, supplied by something beyond thought. To paraphrase Kant, thought would be empty and knowledge would be limited to empirical observation of objects and the usage of language. Yet, even to know that this predicament holds universally and necessarily would transcend the limits of experience and a reason left impotent by being assimilated to mechanical thinking.[4]

The problematic dichotomy lies in the reduction of both mind and machine to either immaterial or inanimate matter, thus Hegel's account concerning machines and mechanisms will provide insights. The presence of AI not only problematizes the notion of the advancement of machines and automation, but the relationality between the human mind and Hegel's Nature, or the externality which it presupposes. Now that we have briefly outlined Hegel's understanding of machines and mechanism, and its relation to the theorizing of artificial intelligence, we can now turn to Marx

[2] Winfield Dien Richard, Hegel, Mind, and Mechanism: Why Machines Have No Psyche, Consciousness, or Intelligence, Bulletin of the Hegel Society of Great Britain 59/60 (London, 2009), pp. 1-18

[3] Ibid., p.4

[4] Ibid., p. 10

and his fragmentary writings on machinery and their effects on capital. Where Marx will follow Hegel is by recognising like Hegel that the machinery appears as an advancement on living labour, but then it begins to deprive labour of its value.[5]

> 'It is clear, therefore, that the worker cannot become rich in this exchange, since, in exchange for his labouring capacity as a fixed, available magnitude, he surrenders its creative power, […]. Rather, he necessarily impoverishes himself, as we shall see further on, because the creative power of his labour establishes itself as the power of capital, as an alien power confronting him ... Thus all the progress of civilization, or in other words every increase in the powers of social production, . . . in the productive powers of labour itself - such as results from science, inventions, division and combination of labour, improved means of communication, creation of the world market, machinery etc. - enriches not the worker, but rather capital; hence it only magnifies again the power dominating over labour; increases only the productive power of capital. Since capital is the antithesis of the worker, this merely increases the objective power standing over labour'.[6]

Marx here clarifies in the advent of machinery the labour that once produced capital then becomes an 'alien power confronting him', in which the inventions of machines do not 'enrich the work but rather capital'. Therefore, Marx sees the presence of automation as a double edged sword as Hegel does as a possible emancipation of the worker from his labour, but as a tertiary object or a technical object machines then drain the previous stage of the conglomeration of the workers labour as a result.[7]

> A word must also be said here, in passing, about the justly famous passages on machinery and automation, which have been so often quoted. Marx here points out, among other things (and, incidentally, this insight is already in Hegel), that with the advance of the division of labour and the growing scale of capitalist production, the role of the worker in the industrial process has a tendency to be transformed from active to passive, from master to cog, and even from participant to observer, as the system of machinery becomes more automatic. Do these passages imply, as some writers have thought, that manual, industrial work, and hence the class which does it, will therefore, under capitalism, disappear, to be replaced, perhaps, by a ' new vanguard ' of engineers and technicians? Such a reading of these passages would be altogether false. It would ignore Marx's unambiguous statements, in many other passages, that there are counter-tendencies which prevent mechanization and automation from advancing beyond a certain limited point, under capitalism ; such a counter-tendency, for example, is the decline in the rate of profit which results from increased investment in machinery relative to living labour. Even in the very same passage on machinery, Marx adds, significantly, that (under capitalism) ' the most developed machinery thus forces the worker to work longer than the savage does, or than he himself did with the simplest, crudest tools '. Neither here nor anywhere else in Marx's work is there a prediction that manual industrial labour will be abolished in capitalist society ; indeed , the weight of Marx's argument carries in the contrary direction.[8]

Therefore, the machine is once produced at a moment by which to alleviate the worker from his labour, but because of the objectivity of capital over the worker forces his further alienation from his labour itself where he becomes simply a 'conscious linkage' in the cogs and organs of the machine itself.[9] What I wish to suggest in examining Hegel's writing on machines and Marx's fragments on automation and machinery, is that in the advent of AI I want to emphasize the far more complex and problematic nature of this advancement. Whilst automata and some forms of machines have existed long before Marx's time, his analysis of the alienating effects of these technical objects is key in understanding how AI as a further dialecticalization of this process of Hegelian Geist could be problematized, where the once inert, inanimate cogs and organs of the machine begin to beat with a heart and mind similar and different to our own.

[5] Marx Karl, Foreword (fore. Trans. by Martin Nicolaus) in Grundrisse, Foundations of the Critique of Political Economy (Rough Draft) Penguin Books, New Left Review, (London, 1973), p. 16

[6] Ibid., Foreword, p. 22

[7] Ibid., Notebook III The Chapter on Capital, p. 308

[8] Ibid., Surplus Value and Profit pp. 376-98 Machinery p. 389 pp. 51-52

[9] Ibid., Notebook VI, p. 692

> No longer does the worker insert a modified natural thing [Naturgegenstand] as middle link between the object [Objekt] and himself; rather, he inserts the process of nature, transformed into an industrial process, as a means between himself and inorganic nature, mastering it. He steps to the side of the production process instead of being its chief actor. In this transformation, it is neither the direct human labour he himself performs, nor the time during which he works, but rather the appropriation of his own general productive power, his understanding of nature and his mastery over it by virtue of his presence as a social body - it is, in a word, the development of the social individual which appears as the great foundation-stone of production and of wealth. The theft of alien labour time, on which the present wealth is based, appears a miserable foundation in face of this new one, created by large-scale industry itself. […] Nature builds no machines, no locomotives, railways, electric telegraphs, self-acting mules etc. These are products of human industry; natural material transformed into organs of the human will over nature, or of human participation in nature. They are organs of the human brain, created by the hum an hand; the power of knowledge, objectified.[10]

In essence, Marx's account of machines and automation even goes as far to set up the relationality of nature, the worker and the object of technics itself. Marx says the worker 'steps to the side of the production process' alongside the technical object which could be analysed in the work of Simondon and Stiegler in placing the technical object not alongside but in between and constitutively constructed. Winfield concludes on this aspect of how the machine could enmesh itself with the living organism below in a succinct way.

> Whether cyborgs, entities that combine machines with a living organism, can qualify as persons is another matter, which can only be settled by exploring the connection of mind and life. That connection is a key problem for the philosophy of mind, long recognised as such by thinkers from Aristotle to Hegel to Searle. How life and mind are related can only be duly confronted, however, once the limits of artificial intelligence have been exposed and the temptation to reduce mind to machine has been repudiated.[11]

The next part shall explore briefly Malabou's work on the brain in her Hegelian reading of plasticity. Malabou below in her conclusion of *What Should We Do With Our Brain? (2008)* summarizes the Hegelian significance and discovery of the relation of the mind to nature.

> This biological alter-globalism is clearly dialectical, as I have said. It demands that we renew the dialogue, in one way of another, with thinkers like Hegel, who is the first philosopher have made the word plasticity *into* a concept, and who developed a theory of the relations between nature and mind that .is conflictual and contradictory in its essence. Rereading his *Philosophy of Nature* could teach us much about the transition from the biological to the spiritual, about the way the mind is real!*y* already a "self *[Setbst],"* a "spirit-nature" at whose core "differences are one and all physical and psychical."' Of course, although Hegel could not yet express himself in the idiom of the "neuronal" and mental his constant preoccupation was the transformation of the mind's natural existence (the brain, which he still calls the "natural soul") into its historical' and speculative"e being. Bur this transformation is the dialectic itself. If there can be a transition from nature to thought, this is because the nature of thought contradicts itself. Thus the transition from a purely biological entity to a mental entity takes place in the struggle of the one against the other, producing the truth of their relation. Thought is therefore nothing but nature, but a negated nature, marked by its own difference from itself. The world is not the calm prolonging of the biological. The mental is not the wise appendix of the neuronal.[12]

Therefore, if the mind is not reducible to the mental nor neuronal in the brain, it can be inscribed elsewhere, and perhaps technics in Simondon and Stiegler is the answer. The next part will be a brief sketch of Simondon's work on Technics as a means by which to understand individuation, or the possible effects of AI on the human being.

> It is possible to read all of Simondon's work as a call for a transmutation in how we approach being. Pursued across physical, biological, psychosocial, and technological domains, this exploration of being assumes a "reformation of our understanding," especially of our

[10] Ibid., p. 705

[11] Ibid., p. 14

[12] Malabou Catherine (fore. Marc Jeannerod), (trans. Sebastian Rand), Conclusion: Toward a Biological Alter-globalism in What Should We Do With Our Brain? Fordham University Press, (New York, 2008), pp. 80-81

philosophical understanding. Expounded in detail in the introduction to L'individu et sa genese physico-biologique, the gesture authorizing Simondon's reflection as a whole receives a definitive formulation at the end of the introduction. Simondon explains that being is used in two senses, which are generally confused. On the one hand, "being is being as such," which is to say, there is being, about which we can initially only confirm its "givenness."[13]

Abbinnett prefaces his work on Stiegler with a quote from Stiegler himself from "How I Became a Philosopher": “I must say that I remain a materialist in the sense of a materialism that does not deny the spirit, but which poses that spirit, while not reducible to matter, is always conditioned by it”.[14] In conclusion, it is apparent that AI demands a new form of ontology which builds upon the understandings of machines and automata, but one that recognises the absolutely differential manner AI will change the relationality between the human mind, its nature and the technical object itself.

References

1 Marx Karl, Introduction, Late August - Mid-September 1857 (1) PRODUCTION Independent Individuals. Eighteenth-century Ideas in (fore. Trans. by Martin Nicolaus) in Grundrisse, Foundations of the Critique of Political Economy (Rough Draft) Penguin Books, New Left Review, (London, 1973), p. 111

Winfield Dien Richard, Hegel, Mind, and Mechanism: Why Machines Have No Psyche, Consciousness, or Intelligence, Bulletin of the Hegel Society of Great Britain 59/60 (London, 2009), pp. 1-18

Malabou Catherine (fore. Marc Jeannerod), (trans. Sebastian Rand), Conclusion: Toward a Biological Alter-globalism in What Should We Do With Our Brain? Fordham University Press, (New York, 2008), pp. 80-81

Combes Murierl, (trans. Thomas LaMarre), On Being and the Status of the One: From the Relativity of the Real to the Reality of Relation in Gilbert Simondon and the Philosophy of the Transindividual, MIT Press, (Massachusetts, 2013), p. 1

Abbinnett Ross, The Thought of Bernard Stiegler, Capitalism, Technology and the Politics of Spirit, Routledge, (London, 2017), p. 1

[13] Combes Murierl, (trans. Thomas LaMarre), On Being and the Status of the One: From the Relativity of the Real to the Reality of Relation in Gilbert Simondon and the Philosophy of the Transindividual, MIT Press, (Massachusetts, 2013), p. 1

[14] Abbinnett Ross, The Thought of Bernard Stiegler, Capitalism, Technology and the Politics of Spirit, Routledge, (London, 2017), p. 1

Acting robots or ethical machines?

Claudia Stancati[1] **and Giusy Gallo**[2]

Abstract. The recent technological developments in robotics and AI bring a greater sense of urgency to the ethical dimension of the future relationships between persons and social robots. To define this relationship, we claim that there is the need to reflect upon the concept of action. First, we will describe von Wright's account of non-causal theory of human action, then we will focus on the actions performed by the so-called social robots. We will point out that the actions of social robots are always planned and predictable while human actions are characterized by creativity. Moreover, spontaneous human actions deal with the spontaneous origin and creation of social institutions.

Keywords. Action, collective action, intentionality, institution, social robot.

1 INTRODUCTION

The recent tendency of some philosophical perspectives considers the main issues on themes such as mind and knowledge stressing their separation from science. On the contrary, some philosophical classical themes have been shaped by changed scientific conditions: for example, the current research in the field of brain computing confirms such assertion so that philosophy could be 'rewritten' starting from this comparison.

The question addressed by the mind-body problem can be read again involving the neuroscientific research, including the provocative theory of the extended mind and the boundaries of the self.

A second issue concerns the nature of learning with reference to the current researches in the field of machine learning.

A third set of questions, related to the other mentioned above, is about the subject of action in relation to actions performed by non-natural subjects and the possible impact, in long term, on the conditions of human sociality

Finally, the philosophical theme of teleology, which has been widely debated in cybernetics, crosses all the issues above mentioned.

It can be argued that Simon has addressed the problem that sums up all the issues mentioned above: what does "artificial" mean? According to Simon the artificial life is "genuine life", although "made of different stuff than the life that evolved here on Earth" [1, p. 33]. Taking into account natural and artificial life as genuine does not mean that one is endorsing a well-balanced position in order to elude issues that, according to Minsky [2, 3], concern the effect of the technological development and AI on the life of men, without excluding the ethical matter from our lives.

2 BEFORE THE ELECTRONIC PERSON

More than a year ago the philosophical debate on Artificial Intelligence was focused on the relationship between human beings and robots, pursuing the quest for an ethical framework [4] within which scanning the issue: do we need a new concept of person after robotics and onlife? Surely, this is a provocative question but it has been encouraged by the discussions began in the European Parliament Committee on Legal Affairs which presented a motion for a Commission on Civil Law Rules on Robotics "in order to begin to establish rules and a kind of ethical and legal code to arrange relationships between human beings and Artificial Intelligence artifacts" [6, 4]. Rights, duties and legal liability ascribed to robots were at stake so far as to envisage the concept of "electronic person" without deeply set boundaries of this concept. Since we have already pointed out risks and perils of a potential anthropocentric view on electronic person [6], we should consider the recent debate on ethics in artificial intelligence [4, 7, 8, 9].

Recently Floridi has coined the term "infraethics": "Consider the unprecedented emphasis that ICTs place on crucial phenomena such as accountability, intellectual property right, neutrality, openness, privacy, transparency, and trust. These are probably better understood in terms of a platform or infrastructure of social norms, expectations and rules that is there to facilitate or hinder the moral or immoral behaviour of the agents involved. By placing at the core of our life our informational interactions so significantly, ICTs have uncovered something that, of course, has always been there, but less visibly so in the past: the fact that moral behaviour is also a matter of "ethical infrastructure", or what I have simply called *infraethics*" [4, p. 390].

Kaplan holds a different position: technological developments affect persons' life and deal with issues such a responsibility, limitations of actions, the transformation of the self, the monitoring of the society. Even though Bodei considers similar themes, drawing future with negative effects on society. Although these positions will not cover the whole debate, no one can deny that each of them implicitly or explicitly implies the notion of action, which is a sort of philosophical cornerstone to be investigated despite the social transformations which involve persons in an everyday life arranged by algorithms and the moral behaviour outside and inside our informational interactions.

[1] Dept. of Humanities, Univ. of Calabria, Italy Email: **giusy.gallo@unical.it**.

[2] The authors have equally contributed to the ideas and content of this article. Claudia Stancati is responsible for sections 1 and 5. Giusy Gallo is responsible for section 2, 3, 4.

3 A PHILOSOPHICAL STANCE ON ACTION

Probably, in philosophy, the best-known description of human activity is the Aristotelian distinction between *poiesis* and *praxis*, which has often been recalled by ancient, modern and contemporary philosophers interested in agency and, more generally, in human nature. It has to be pointed out that at the end of Nineteenth century French philosophers began to recall the value of action within the so-called philosophy of action and spiritualism (e.g. Blondel, Bergson), but a few decades later analytic philosophers – Ryle and Wittgenstein before, von Wright and Anscombe later – dealt with the concept of human action and freedom.

Von Wright maintained an original position about the relationship among action, causation and free will: he refused the causal (in the sense of natural sciences) explanation of human action and shaped a notion of determinism which is compatible with the freedom of action in a teleological perspective:

If we could not, *on the whole*, account in terms of reasons for what people do, it would be difficult for us to understand them *qua* agents. If this were the case with ourselves, we should cease to feel responsible for our actions, since we could not then on the whole account for them. We should perhaps think that we are at the mercy of uncontrollable outer or inner forces - maybe of a causal nature.

Free action and action for reasons are twin concepts. “Determination” of action through reasons is, one could say, a precondition of human freedom. Without this type of determination our very notions of agent and action would not exist, or be quite different from those we have [10, p. 133].

In von Wright’s theory of action, the agent is not always able to self-understand his action also if he can freely choose among actions in terms of reasons and balancing the *future*. The root of determinism is neither the betrayal of free action nor the denial of free will:

Human freedom, it was then often said, just consists in this: that an agent’s actions are determined by his will and not by external forces over which he has no control or power. This was a way of reconciling freedom with determinism (cf. below, 152). It was thought important as long as science nourished and sanctioned a deterministic world-view. But a difficulty was lurking in the background. Granted that action is free when in conformity with our will, what then of the will itself? Are we free to will what we will? Or is the will determined by something else? If the will is not free, action determined by the will can be free at most in some relative sense, it seems. Willing has an object, is *of* something. And the same holds for intending, wanting, and wishing. Only seldom do we explain an action by saying that we willed or wanted just it. Giving this answer is more like brushing the question of why we did it aside — like saying “it is none of your business to inquire into the motives for my action.” The reason why I did something might be that I coveted or wanted something else to which I thought the action conducive. This other thing was then the object of my will. Willing *it* was the *reason* for my action, that which made me do what I did [11, p. 2].

Willing does not cause action but the *object* of willing has a causal role in performing an action. In the sphere of willing and free action, von Wright rapidly shifted to the concept of person in order to define the background of reasons, in which he pointed out the social aspects of human life:

In attributing reasons for action to an agent we normally also attribute to him various abilities, beliefs, desires and inclinations, the understanding of institutions and practices of the community, and other things which characterize him as a person. Some of these features may date far back in his life history. They constitute a kind of background or “program” which has to be assumed if certain things he did or which happened to him shall count as reasons for subsequent action (for example, that he understands a certain language). These other things, then, speaking metaphorically, are “inputs” playing on the “keyboard” of his programmed personality. His action is the “output” [11, p. 27].

Even if the comparison to an input/output machine seems to be crucial for thinkers who are trying to widen an ethical perspective in Artificial Intelligence research, we will focus on the first part of the above quote: the agent is a person whether he can be recognized as involved in (at least?) institutions and practices of a community. In fact, we might consider that understanding an institution or understanding a practice means to reduce them to the sum of individual actions of persons with their biological equipment and intentionality. This idea shows a weak point: human agency is confined in a chain of micro-actions which fall under a mysterious law. How to explain the relationship between individual action and the collective institutional action? Von Wright’s perspective does not claim for such issue. The reference to human social attitude claims for a subsequent philosophical question: how does an institution or community act?

4 WITHOUT BEING AFRAID OF SOCIAL ROBOTS

The question about the agency of institutions and practices could be arranged in reference to the agency of social robots.

The social robot is an autonomous or semi-autonomous robot that is involved in some programmed and scheduled practices which are structured taking into account the role they play in interactions with human beings and/or other social robots.

Nowadays, robotics offers two kinds of social robots which are subject of debate among philosophers, AI scientists, sociologists and politicians: care robot for elderly persons and butler robot.

The care robot is programmed to help elderly persons in doing a lot of things: going upstairs and downstairs, (making sure of) taking medicines and calling the doctor, checking the kitchen, interacting and communicating. It seems that care robots can help elderly persons in their everyday life but psychologists and sociologists glimpse some risks: does this kind of social relation meet the needs of elderly people? Is there the risk of an emotional attitude towards the care robot? How to protect and promote elderly persons’ social life?

The butler robot acts like a servant: it helps in housekeeping, it monitors the house when the owners are gone, it watches on children while they are playing or doing homework, it keeps an eye on the pets, it interacts with all the persons the robots meets in the house.

Then, we could imagine a near future in which we can choose to improve our quality of life by including in our family life a social robot. Someone might consider this idea as a provocative one but this is not science fiction. Different categories of robots are already in our lives so that there is no reason to consider social robots dangerous although there are some legal and ethical issues to be examined. One critical juncture is the chance of disruptions and malfunctions of the social robot and eventual damages to the persons he has to take care of. Who will be responsible for those damages? Probably it depends on the context in which the damage occurs.

The reference to this scenario is useful in order to clarify the concept of action performed by social robots which act as artificial companions [9]. The range of actions of social robots is planned and ruled by software which activates under certain conditions and releases as output an action (including *verbal behaviour*) compatible with his role.

Does an artificial companion act as a human being? Yes, whether we consider only the "output". No, whether we aim to consider an artificial companion as replacement of human beings, which means to imagine an apocalyptic future of superintelligent machines which can rule over their makers and designers. Despite the great enhancements in robotics and AI, robots lack of the kind of creativity ascribed to human beings (even if machines are creative and can learn in other ways). Moreover, all actions are predictable (unless malfunctions), formalized and goal-directed, even when the machine has to 'calculate' the best action to perform in order to solve a problem following rules (including social customs, if they are scheduled).

The philosophical theory of action mentioned in the previous section does not cover the actions of a machine since it represents a dynamic idea of human action based on temporality while the range of actions performed by a robot is static unless the software undergoes an (automatic) update, which has to be planned by an ICT expert. Not unless an external intervention, the social robot is not capable of actions that need cooperation and innovation. For example, if we consider the communicative powers of social robots we should ask how Jarvis, the butler robot, can understand a baby who is learning the first words or an elderly person with speech disorder who is asking for a glass of water. Does Jarvis understand the slang spoken by the teenagers he is monitoring while the parents are gardening?

The examples mentioned above show that the work on social robots has only begun but also in the case of great improvements, the nature of robots will probably always collide with the tacit nature of human practice and institution, enclosed in the following lines: "persons mutually adjust their full-time activities over a prolonged period, resulting in a complex and yet highly adaptable co-ordination of these actions" [12, p. 141].

5 COLLECTIVE ACTIONS AND INSTITUTIONS

Considering the philosophical category of action in light of AI shows a fundamental problem which involves the categories of act, consciousness, intentionality and representation.

Between the end of 19th century and the beginning of 20th century the notion of intentionality has been discussed and analyzed by Brentano and Meinong culminating in the deep and sophisticated analysis of Husserl, despite in his thought there is not a univocal formulation of that concept. But if, following the authors mentioned above, we decouple the concept of act and the concept of activity, the first one deals with the mental and / or intentional sphere and the second one deals with the agent as a device or a vector.

During AISB 2017 convention, we have pointed out that the legal personality could be useful in order to define the personality of robots: in particular, legal personality is a tool which legitimates the completion of acts and activities and it allows the allocation of responsibility in relation to the consequences of action.

We would like to steer the reflection from individual action to collective action in a very peculiar sense, without any pretense in giving solutions. Our idea is not about thinking a collective subject as a singular subject, such as an entity with legal personality, but we would like to reflect upon the consequences of social agency in the light of the new "centers of action" which we hesitate to call subjects even if they act, learn and interact also with human beings.

Collective action, the nature of collective subjects and the relation between the so-called *macro* and *micro* level are the heart of the sociological debate. However, in all the fields in which the debate deals with institutions, the main issue refers to intentionality in the sense of voluntary and planned human actions which lead to the constitution of institutions.

In sociology there are two main approaches about the origin and the development of institutions. On one hand, one approach shows that individuals are the only existent reality and the institutions are the unexpected and unintentional result of their action. This is a non finalist perspective on institutions. For example, Menger, who endorses the Scottish thinkers of the late Enlightenment, distinguishes institutions in organic unintentional institutions and organic and pragmatic institutions, also defining a kind of mixed institutions. Those positions are shared by Simmel and Tarde, so that there is a sort of division of labour on a historical basis then everyone can use acquired and accumulated results from older generations. According to Hayek, institutions are the result of human action, which is not planned, following a spontaneous but not artificial order.

On the other hand, there are thinkers following Durkheim for whom social fact are more than the sum of individual actions, but social objects are equipped with proper willingness and reality which is separated from the reality of the elements that compose actions. The opposition between methodological individualism and holism, both declined in various forms, is still one of the crucial knots of social sciences.

One of the most famous ways to think the collective dimension is the we-mode, quoting as example Raimo Tuomela:

> The we-mode approach is based on the intuitive idea that the acting agent in central group contexts is the group viewed as an agent, and the individual agent is not the primary actor but rather a representative acting for the group. To go into some detail, according to our intuitive view the group can be regarded as an agent from a conceptual and justificatory point of view, but ontologically it exists only as a social system, not as an agent, and it can only function through its members' functioning appropriately. The ontological and causal work is done by the members' actions and joint actions and what these produce. My conceptually and psychologically holistic starting point is simply that there is a group (an instrumentally viewed agent) that is the intentional—but not the ontological—subject of attitudes and

actions attributable to it. The group, which is assumed to be a "we-mode group" in my terminology, is assumed to commit itself to a group "ethos" (certain constitutive goals, beliefs, standards, norms, etc.) and to acting accordingly. This intuitive picture can be explicated for the group-member level and seen to involve three central ideas to be called "authoritative group reason," the "collectivity condition," and "collective commitment" (see below). In contrast, in the I-mode case the individual is the sole acting agent. This is a crucial difference, which my we-mode approach tries to make sufficiently clear and ontologically palatable. Translated into the group-member level the above holistic view gives this: The group members function as group members as if they were cogs in a machine, viz. the group agent capable of action. Because of this, the group's ethos, as a central "jointness" element assumed to be (extensively) accepted by the group members, gives them their central reason for acting as group members. The reason is an authoritative one if the group members themselves have participated in the creation of the ethos by their collective acceptance. Similarly, because of being members of a group (qua agent), the group members will necessarily "be in the same boat" when acting as group members. This will be explicated by the collectivity condition, the satisfaction of which comes about through the members' collective commitment to the ethos and action on the basis of this commitment. The we-mode group's commitment to its ethos basically amounts to the members' collective commitment to it. Here the conceptual starting point is the group's accepting an ethos with commitment to its satisfaction and maintenance. On the group-member level, this amounts to the group members' performative collective acceptance (indeed, collective construction) of an ethos (e.g. a goal) as the group's ethos (goal) to which they collectively commit themselves, where collective commitment accordingly is "reasoned" by the group commitment and where collective commitment also involves the members' being directedly socially committed to each other to functioning as group members, typically to furthering and maintaining the ethos. We-mode thinking, "emoting", and acting accordingly presuppose reflexive collective acceptance ("construction") of the group's ethos and often also of some other, nonconstitutive content as the object of the group's attitudes. The collectively accepted contents must be taken to be for the benefit and use of the group. In all, the members' are taken to view and "construct" their (we-mode) group as an entity guiding their lives when their group membership is salient, and it also requires them to function as ethos-obeying and ethos-furthering group members thus as "one agent". To recapitulate, the we-mode is taken to involve the notion of a social group in a strong sense involving the strongly interconnected features of an *(authoritative) group reason*, *collective commitment* of the members to the group's reason-giving ethos (reflecting, as we may say, the group's commitment to its ethos) and the resulting group uniformity making the *collectivity condition* satisfied on the member level [13].

According to us, in this quote there is a paradox: while the tendency of our thought is to consider the other, that is an animal, a robot or an alien from a different world, in an anthromorphic view, when we think about the collective action, the mechanical metaphor gets over, as Tuomela shows, in all the thinkers who are not able to conceive institutions as spontaneous and not planned, namely thinkers who need the metaphor of artificial to describe the state and the society.
Descartes, writing to his friend Chanut, refusing the anthropocentric vice, allowed to other hypothetical intelligent beings which could live in his universe all the things he denied to the animals with which we share the earth:

In the same way I don't see that the mystery of the Incarnation, and all the other favours God has done to man, rule out his having done countless other great favours to countless other creatures.
I don't infer from this that there are intelligent creatures in the stars or elsewhere, but I don't see any argument to show that there aren't. I always leave questions of this kind undecided, rather than denying or asserting anything about them. The only remaining difficulty, I think, is that we have long believed that man has great advantages over other creatures, and it looks as if we lose them all when we change our opinion about what thinking beings there are in the universe, the fear being that if there are countlessly many of them on other planets we may lose all our privileges because we're outranked by them. I now allay that fear […] Now the goods that could belong to all the intelligent creatures in an indefinitely large world belong to class (b); they don't diminish our goods [14, pp. 200-201].

REFERENCES

[1] H. Simon. *The Sciences of Artificial*, Cambridge MA, MIT Press (1981).

[2] M. Minsky. Alienable Rights. In *Android Epistemology*, Cambridge MA, MIT Press (1995).

[3] M. Minsky. *The Emotion Machine. Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind*, New York, Simon&Schuster, Cambridge MA, MIT Press (2006).

[4] L. Floridi. *Infraethics – On the conditions of Possibility of Morality*. *Philosophy and Technology, 30* (2017)

[5] Committee on Legal Affairs (Rapporteur Mady Delvaux). European Parliament resolution with recommendations to the Commission on Civil Law Rules on Robotics: http://www.europarl.europa.eu/committees/en/juri/subject-files.html?id=20170202CDT01121.

[6] G. Gallo and C. Stancati. Persons, Robots and Responsibility. How an Electronic Personality Matters. In: *AISB Conference Proceedings*, Bath (2017).

[7] N. Bostrom. The Ethics of Artificial Intelligence. In: *Cambridge Handbook of Artificial Intelligence*, Cambridge UK, Cambridge University Press (2011)

[8] N. Bostrom. *Superintelligence. Paths, Dangers, Strategies*. Oxford UK, Oxford University Press (2014).

[9] L. Floridi. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford UK, Oxford University Press (2014).

[10] H. Von Wright. Explanation and Understanding of Action. *Revue Internationale de Philosophie*, 35 (1981).

[11] H. Von Wright. Of Human Freedom. In: *In the Shadow of Descartes: Essays in the Philosophy of Mind*, Springer, The Netherlands: (1998).

[12] M. Polanyi. *The Logic of Liberty*. London UK, Routledge and Kegan Paul (1951).

[13] R. Tuomela. *Group Thinking*. Invited paper for Collective Intentionality VI Conference, UC Berkeley, USA (2008).

[14] R. Descartes. *Selected Correspondence*. J. Bennet (2013).

From the Chinese Room Argument to the Church-Turing Thesis

Dean Petters [1] and **Achim Jung** [2]

Abstract. Searle's Chinese Room thought experiment incorporates a number of assumptions about the role and nature of programs within the computational theory of mind. Two assumptions are analysed in this paper. One is concerned with how interactive we should expect programs to be for a complex cognitive system to be interpreted as having understanding about its environment and its own inner processes. The second is about how self-reflective programs might analyse their own processes. In particular, how self-reflection, and a high level of interactivity with the environment and other intelligent agents in the environment, may give rise to understanding in artificial cognitive systems. A further contribution that this paper makes is to demonstrate that the Church-Turing Thesis does not apply to interactive systems, and to self-reflective systems that incorporate interactivity. This is an important finding because it means that claims about interactive and self-reflective systems need to be considered on a case by case basis rather than using lessons from relatively simple non-interactive and non-reflective computational models to generalise to all computational processes.

1 Introduction

This paper will show that Searle's Chinese Room Argument (CRA) scenario [14] can be extended and given more detail so that new variations of this scenario have a fundamentally different relationship with the Church-Turing Thesis (CTT). Searle's CRA is a gedanken experiment aimed at demonstrating that computer programs cannot really understand the meaning of what they process, even if their observable behaviour seems to demonstrate understanding. The CTT is commonly interpreted as stating that the intuitive concept of computability is fully captured by Turing machines or any equivalent formalism (such as recursive functions, the lamba calculus, Post production rules, and many others). The CTT implies that if a function is (intuitively) computable, then it can be computed by a Turing machine. Conversely, if a Turing machine cannot compute a function, it is not computable by any mechanism whatsoever.

This paper presents a family of variations to the CRA which involve changing the CRA to require significantly more interaction with the outside world: in frequency of interruptions; interleaving of interruptions; and in the nature of the information provided by interruptions. A second family of variations to the CRA includes the same pattern of interruptions and close coupled interaction with the external environment but also includes ways in which higher-level routines within the CRA program can analyse the basic program for 'meaningful' patterns in its own internal processing. These versions of the CRA are outside the scope of the CTT because the CTT is concerned with situations where programs act as mathematical functions with inputs fully provided at the start of the computation and with no possibility of new inputs being included during the run of the program. This matters because the CTT is commonly invoked to generalise the lessons from the CRA to *all forms of computation whatsoever* while it is only legitimate to draw conclusions about programs and computational mechanisms which follow the basic input-output paradigm. If programs presented in new variants of the CRA scenario fall outside the scope of the CTT (but are still recognisable and implementable programs in the sense of being precisely specifiable algorithms) then the lessons from these variants will not necessarily generalise to all possible programs. Therefore, any generalisations would need to be validated on a case-by-case basis for prospective program formalisms. The paper concludes with the observation that the new variant CRA scenarios sketched in this paper are not only more similar to typical human cognition than the very simplified portrayal of processing in the original CRA, but the complexity they present is fast being achieved and overtaken by contemporary computing systems.

2 Overview of the CRA — and how lessons drawn from it are generalised

Published in 1980 in the paper *"Minds, Brains, and Programs"*, [14], Searle made an argument based on a 'Chinese Room'. It is a thought experiment that is intended to show that running programs cannot have understanding and awareness of what they are doing. Searle introduced his first CRA scenario by discussing an earlier simulation produced by Roger Schank and co-workers, [13]. Searle explained that he was using that work as inspiration for this CRA scenario because of his own familiarity with this program. However, he also claimed that his argument does not rely on the details of Schank's programs, and in fact applies to any Turing machine simulation that is modelling mental processes. It is this claim of generalisation to all programs (because Schank's program can be run on a Turing machine) that is the critical focus of the present paper.

Schank's program simulates the ability to understand stories. The program accesses information about particular contexts and the program can then answer questions about a story set in that context. This is accomplished by analysing what is stated in the story and what can be expected in the context in which this particular story is set. Schank's program accomplishes this by possessing a representation, which he terms a 'script' that includes contextual information of the sort that humans possess. Searle highlights the fact that in this process it is only the form of the representations of the story and script that are necessary and sufficient to produce the output. The content of the representations of the story and script takes no part in the algo-

[1] University of Wolverhampton, UK, email: d.petters@wlv.ac.uk
[2] University of Birmingham, UK, email: A.Jung@cs.bham.ac.uk

rithmic process and is not required to transform input to output. The key lesson that Searle draws from the CRA is that the formal symbol manipulations carried out within the Chinese room do not give rise to meaning or understanding, operations in the Chinese room are all *"syntax but not semantics"* ([14], p. 422).

3 Different varieties of Searle's Chinese Room scenario have fundamentally different relationships with the CTT

3.1 Searle's original scenario

> *'Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles.*
>
> *Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch 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. Unknown to me, 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." '* ([14], p. 418)

Searle's lesson is that an observer external to the room would see meaningful behaviour but within the room there is only meaningless symbol processing — so demonstrating that understanding cannot arise from just the operation of formal syntactic processes.

We can see that this very abstract description of a running program is not only based on a single run of the Schank program, but also matches the classic modus operandi of the very early generations of electronic computers. These carried out 'batch jobs' where the input and program were both completely specified at the start. The computing machine would process the input according to the program, and the output would appear as a paper printout. Contemporary computing no longer works like this, with many possible interruptions to ongoing processing. The next scenario attempts to sketch out different ways in which interruptions and new input data can appear during the running of a program.

3.2 Searle managing multiple tasks by effectively processing real-time updates from the environment

This quote from Monsell highlights the delicate balancing act in natural systems between forcing through ongoing processing on a primary task and dealing appropriately with potential interruptions:

> *"Hence the cognitive task we perform at each moment, and the efficacy with which we perform it, results from a complex interplay of deliberate intentions that are governed by goals ('endogenous' control) and the availability, frequency and recency of the alternative tasks afforded by the stimulus and its context ('exogenous' influences). Effective cognition requires a delicate, 'just-enough' calibration of endogenous control that is sufficient to protect an ongoing task from disruption (e.g. not looking up at every movement in the visual field), but does not compromise the flexibility that allows the rapid execution of other tasks when appropriate (e.g. when the moving object is a sabre-toothed tiger)."* ([10], p. 134).

Following Monsell, an interactive CRA scenario could capture the closely coupled nature of interactions between agent and environment and might involve running many sub-programs in parallel with an overarching program acting as a kind of operating system. Not only does Searle's 1980 scenario completely ignore the nature of algorithms that require this level of constant checking the current state of the environment, it also ignores the nature of 'forever' processes such as operating systems carrying out processes such as resource management and process control, but then returning to the same ground state and never providing a final output. Yet it seems something like this kind of 'operating system' algorithm must be implemented in humans and other animals. In addition, the ever increasing complexity of artificial control systems like intelligent mobile robots and self-driving cars can increasingly be seen to incorporate these kinds of complex interactions, driven by environment interruptions alongside the requirements of multiple primary tasks. So adding further complexity, interactive variants of the CRA scenario might also include parallel computation, probabilistic computation, and real-time computation, all of which are manifestly outside the scope of the CTT.

This leads to a key claim of this paper — an interactive variant of the Chinese Room Argument scenario similar to Monsell's description, getting input during the running of the program, but also including multi-processing, real-time computing, (truly) probabilistic computation, programs that never terminate, distributed computation, intentional computation, and higher order computation, is outside the scope of the CTT. This is because the CTT is concerned only with the equivalence of systems that operate from input provided as a string to output as a string; it does not cover programs that deal with a series of inputs, appearing over time and at unexpected moments, or that are capable of making changes to the operation of the program while it is running. As laid out clearly and carefully in his 1936 paper, Turing's concern was with the steps a mathematician (a human "computer" in Turing's terminology) goes through when following a precisely and finitely specified procedure. His analysis is compelling and we see no reason not to accept the CTT, though we emphasise its constrained setting. The (partial) functions from strings to strings that can be computed by Turing's "machines" are the same ones that can be computed by any and all formalisms that have to date been put forward as alternatives. Put another way, the evidence for the CTT is very strong indeed, but this does not give licence to applying it — by analogy, as it were — to other forms of computation. These "other forms of computation" are not the fruits of idle speculation but very much day-to-day reality for software engineers and computer users alike: computing machines no longer expect a question on a tape (or punch cards), go away and compute, and return an answer as a single file (or more punch cards or some output paper). Contemporary programs interact with their environment in multiple ways, and employ facilities (such as hardware-based random number generators) that can not be shoehorned into the paradigm for which the CTT was formulated.

4 The CTT in theoretical computer science and its implications for the philosophy of mind

4.1 Functions versus processes

The misconception that the CTT applies to all forms of computation is very wide-spread also within the computer science community, and even among theoretical computer scientists. Goldin and Wegner, [6], examine the likely origins of this belief, which they term the 'Strong Church-Turing Thesis', and explore the reasons why it holds such sway. They suggest that the way the first generation of computing machines were designed and used (i.e., the "batch processing" discussed above) was so strongly correlated with Turing's mathematical concept of a (human) "computer" (i.e., his "Turing machines") that standard undergraduate textbooks adopted Turing machines as a suitable formal abstraction of computing practice. Like us, Goldin and Wegner point out the role of interactivity that is so central to modern computing systems, and that is simply not covered by the CTT.

Some researchers have been very aware that Turing machines are not appropriate for modelling interactive behaviour and have proposed alternative mathematical abstractions. We mention especially the work of Milner, [9], and Hoare, [8], on computational "processes". It is astonishing (but not the focus of the present paper) that although their work has been incorporated into undergraduate syllabuses for decades, courses on computability theory still promulgate the view that Turing machines are all there is to computation.

Beyond the analysis for this state of affairs given in [6] we believe that it is useful for our argument to point out one crucial difference between the setting of the CTT and the more encompassing computational models of Milner and Hoare: When we consider computation from fixed input to single output (the "function view" of computation), then the equivalence of computational mechanisms is almost unavoidable. To give just one example, it is not the case that the λ-calculus was designed with computability in mind; rather, its purpose originally was to give a new foundation for mathematics, replacing set theory (see [2] for a historical introduction). As far as functions from natural numbers to natural numbers are concerned, the equivalence with Turing machine computability was noted *afterwards*. In contrast, mathematical models for interactive behaviour (the "process view" of computation) can be *quite different* in expressivity. A canonical, maximally expressive formalism for processes simply does not exist. We point the interested reader to Abramsky's [1] where this fact is highlighted and explored.

One final comment on the difference between the functional and the process point of view: It is, of course, possible to use a rich interactive machine to implement a simple function; after all, that is what we do with our modern computers all the time. It is our belief that this does not lead to new computable functions, i.e., some sort of "hypercomputation". In other words, the CTT is valid even if more sophisticated machinery is employed. It is the other direction that is the core of this paper: When considering more sophisticated computational tasks, then standard Turing machines (and their mode of operation) are not sufficient to explore the range of possibilities.

4.2 Computation in an extended sense

So far, we have focused on interactivity as a (ubiquitous) feature of modern computational systems which is not present in the Turing machine model. There are others which are also interesting for our argument, especially in an interactive setting. We begin with the question whether the computational process has internal memory or not. If it does, then it can react differently to identical stimuli from the environment as time passes, and indeed it can exhibit "learning behaviour". A study of this facility from the point of view of computability theory is presented in [5], for example. What is important for our argument is the fact that an interactive process that has some finite internal memory is strictly more powerful than a process that does not, and a process that has unlimited internal memory is strictly more powerful than one with finite memory. Thus we have a fairly straightforward computational situation were the CTT is false, or to be more precise, where there is no analogue of the CTT.

If we translate these findings to Searle's Chinese Room, then we are in a situation where he may be in interaction with his environment, constantly receiving and issuing statements expressed in Chinese characters. Having the ability to make personal notes (in English but perhaps with Chinese characters interspersed) would greatly enrich his experience and might even lead him to understand the meaning of these interactions. This would be true even if the form of his note-taking were already prescribed in his original "script".

A similar argument can be made for processes that have access to a real-time clock, or to a source of true random numbers.

4.3 Implications for the philosophy of mind

Both Goldin and Wegner, and Abramsky, highlight an issue in theoretical science which has not yet been received in the philosophy of mind literature. They show that for computer scientists the CTT should be treated as a thesis that certain models of computation are equivalent for tasks that require the transformation of given fixed finite input to some output, and not that TMs can implement every possible kind of information processing machine. Goldin and Wegner do argue that researchers in Artificial Intelligence are somewhat ahead of researchers in theoretical computer science in promoting interaction rather than computation of functions as beneficial in expressing the behaviour of information processing systems. For example, they cite Rodney Brooks' 1991 statement of interaction as a prerequisite for intelligent system behaviour [3], and Russel and Norvig recognising that intelligent behaviour is better modelled by interactive agents than functions with prestated inputs and outputs that only occur at the termination of the computation [12]. However, this mistaken view, that the CTT states that TMs are capable of such broad information processing capabilities, seems to be what justifies the generalisation that lessons from Searle's specific CRA scenario applies to all possible programs. For example, in *The Critique of Cognitive Reason'* Searle invokes the CTT to state that for any algorithm there is a TM which can implement that algorithm — which is a correct interpretation of the CTT (assuming the common interpretation of "algorithm"). However, he then goes on to suggest that the next step from this line of reasoning is that the brain is a Universal TM. Whilst he concedes that in addition to algorithmic processes (within the scope of the CTT) there may be unconscious processes outside the scope of the CTT, he does not consider that processes which link up and transition between individual computations are of this unconscious type. In fact, he does not consider dynamic and contingent transitions between individual function-based computations at all ([15], p. 837).

4.4 Searle carrying out self-reflection of his own program

New interactive variants of the CRA may be outside the CTT, but they do not necessarily demonstrate more understanding in the inner workings of the Chinese room. Inserted information may be just as impossible for 'Searle in the room' to understand as the Chinese symbols in the original CRA scenario. However, we can not only vary frequency, interleaving, and parallelisation due to interruptions, but also form new CRA scenarios which involve kinds of information that are intended to interact with the running program to change the English rules that Searle carries out. The CTT does not cover processes where in principle any information (from a simple boolean to an analysis of the running of the existing program to a whole new program) can be added as input during the running of the program. In the book *'Kinds of Minds'*, Dennett [4] portrayed a number of different abstract agents (creatures) according to how they processed information. He presented Darwinian creatures as not capable of learning but acting upon evolved reflexes; Skinnerian creatures as learning from association; and Popperian creatures that can pre-select strategies after evaluating their likely success in internal working models. In addition to these creatures, Dennett also described Gregorian creatures, *"whose inner environments are informed by the designed portions of the outer environment."* ([4], p. 99). Thus Gregorian creatures can import 'mind tools' wholesale from the environment ([4], p. 100). What is relevant to the CRA and Searle's conclusion is whether the inputs to the Chinese room can not only add to the store of Chinese symbols but also add to or substitute for the English instructions that 'Searle in the room' actually follows. Programs which can be interrupted to receive new information that may alter in a fundamental way their processing, even conceivably by changing the running program itself, are clearly outside the scope of the CTT. This is because if a new program can be given as input on an interruption, this is no longer the program which started processing.

It is possible to have algorithms in the Chinese Room that engage in self-reflection and self-analysis. When self-reflection and self-analysis occur it can create a kind of internal 'meaning' about the system which may then be linked to external meaning in the form of patterns in Chinese symbols. Any 'Searle in the room' can only carry out the English instructions which are directly given to him. 'Searle in the room' can never do anything which is not part of a task set out in English instructions. But a 'Searle in the room' can follow the programmed instructions for the specific 'narrow' task at hand of processing stories, and his overall task can also involve a whole set of further instructions which may be triggered at any time, and often are triggered by well considered interruptions from the outside, and which involve questioning what the nature of the connections between internal rules and data mean. He can be asking, in addition to what patterns in input and output data exist, what patterns exist in the use of his English rules. When do rules co-occur? What rules predict other rules? Do some rules being triggered predict the task is nearly over? Are some patterns more surprising than others? Are there clusters or categories of rules that perform similar tasks? The 'Searle in the room' accomplishing this broader and self-reflective task is not carrying out a non-algorithmic process. Rather, he is still following English rules that compare rules and processes looking for identifiable patterns. But these patterns do not then trigger the outputting of meaningless symbols. Rather, Searle is learning about meaning in processing patterns apart and aside from the meaning of the symbol tokens being processed. Meaning is emerging from the internal processing of rules set apart from the meaning of the Chinese symbols.

These 'Searle in the room' self-reflection scenarios highlight distinctions between: (i) algorithms that carry out specific narrowly defined tasks, and just carry out those tasks versus algorithms that carry out tasks and simultaneously search for meaning in the properties and implications in patterns in running processes and events; and (ii) 'representational' meaning by understanding the content of symbol tokens versus 'dynamic processing' meaning by understanding the properties and implications in patterns in running processes and events.

Further variants of self-reflection scenarios include: 'Searle in the room' looking for meaning but making no connection to the external meaning of the Chinese symbols, so all meaning emerges from the bottom up; and, 'Searle in the room' looking for meaning internally and connecting this to appropriate external symbols — thus giving external symbols more than derived intentionality. He might be helped in this by people outside the room interacting with him in a way designed to foster the emergence of meaning.

5 Importance of the CTT for psychology and cognitive science

In his review of the literature around the CRA, Preston makes clear why CTT matters to psychology and cognitive science:

> *"Even more important than the nature of the thesis, perhaps, is the matter of its implications. It's no exaggeration to say that the Church-Turing thesis has constituted the fundamental inspiration behind AI, the reason for thinking that electronic digital computers must be capable of (at least) human-level intelligence. Cognitive scientists have generally taken the Church Turing thesis to mean that any function that can be computed can be computed by a Turing machine. This would mean that, as long as we ignore or abstract away from resource limitations, anything the human brain can do (any function it can compute) could also be done (computed) by an electronic digital computer. Cognitive processes, no matter how intelligent must be decomposed into routines whose primitive steps can all be executed by a machine"* ([11], p. 6).

Since it was first formulated in the 1940s no-one has really questioned the CTT, and nor do we. It is one of the jewels of theoretical computer science [7]. However, the CTT is concerned only with functions from from strings to strings — input needs to be given as a fixed finite string and output (if it is produced) will be a finite string. The CTT underlies the strength of the CRA because it allows Searle to say: 'the limitations of the program in the CRA applies to all programs because of the CTT'. Accepting this generalisation strategy, as Searle does, means there can be no syntactic formal processes in a program, of even fiendish complexity or strangeness, that will ever give rise when running to any kind of semantics. If, on the other hand, Searle cannot invoke the CTT for the CRA then whatever lessons he draws from the CRA only apply to the specific scenario he presents.

6 Conclusion

This paper takes the position that there are implementable programs which are outside the scope of the CTT. The central argument of this paper is that invoking a mathematical theorem to make inferences about real-time physically instantiated systems should be done with careful consideration of both the scope of the theorem and the properties and complexity of the physical system. Turing set out to solve the “Entscheidungsproblem” (decision problem) and for this purpose proposed a mathematical formalism that faithfully emulates the process of a human being following finitely specified instructions. It was soon found that other formalisms have the same expressive power in this specific setting, i.e., mathematical problem solving, and this then led to the CTT. Situations in contemporary computing are now so rich, they can no longer be said to be covered by a paradigm where the inputs are known in advance, the system is left alone to do its computation and then provides the answer. Critically, for richer kinds of computation, some of which have been described in this paper, the empirical evidence suggests that there are many shades of expressivity, which is why no-one has ever postulated an analogue of the CTT for them.

This paper therefore agrees with Searle insofar as when programs confirm to the requirements for CTT equivalence, there can be no meaning in the internal symbol processing. But for programs outside the scope of CTT, meaning can appear in several ways, (i), by interaction with the environment, and (ii), by self-reflection within the program. This paper challenges the idea that syntax (in the case of a running program) cannot give rise to semantics. Therefore this paper takes a radical approach which attempts to overturn 80 years of misguided extrapolation that the CTT applies to all programs that are of interest to computer science, cognitive science, and philosophy.

REFERENCES

[1] S. Abramsky, ‘Intensionality, definability and computation’, in *Johan van Benthem on Logic and Information Dynamics. Outstanding Contributions to Logic, vol 5.*, eds., A. Baltag and S. Smets, 121–142, Springer, (2014).

[2] H.P. Barendregt, *The Lambda Calculus: Its Syntax and Semantics*, North-Holland, revised edn., 1984.

[3] R. Brooks, ‘Intelligence without reason’, in *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1*, 569–595, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, (1991).

[4] D.C. Dennett, *Kinds of minds: towards an understanding of consciousness*, Weidenfeld and Nicholson, London, 1996.

[5] D.Q. Goldin, S.A. Smolka, P.C. Attie, and E.L. Sonderegger, ‘Turing machines, transition systems, and interaction’, *Information and Computation*, **194**, 101–128, (2004). Special Issue Commemorating the 50th Birthday Anniversary of Paris C. Kanellakis.

[6] D.Q. Goldin and P. Wegner, ‘The Church-Turing thesis: Breaking the myth’, in *New Computational Paradigms*, eds., S. Barry Cooper, Benedikt Löwe, and Leen Torenvliet, pp. 152–168. Springer Berlin Heidelberg, (2005).

[7] D. Harel and Y. Feldman, *Algorithmics: The Spirit of Computing*, Springer, 3rd edn., 2012.

[8] C.A.R. Hoare, *Communicating Sequential Processes*, Prentice Hall International, 1985.

[9] R. Milner, *A Calculus for Communicating Systems*, volume 92 of *Lecture Notes in Computer Science*, Springer Verlag, 1980.

[10] S. Monsell, ‘Task switching’, *Trends in Cognitive Science*, **7**, 134–140, (2003).

[11] J. Preston, ‘Introduction’, in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, ed., J. Preston, 1–50, Oxford University Press, Oxford, (2003).

[12] S. Russell and P. Norvig, *Artificial Intelligence, A Modern Approach, 2nd Edition*, Prentice Hall, 2003.

[13] R.C. Schank and R. Abelson, *Scripts, plans and understanding*, Lawrence Erlbaum Associates, Hove, UK, 1977.

[14] J.R. Searle, ‘Minds, brains, and programs’, *The Behavioral and Brain Sciences*, **3**(3), (1980). (With commentaries and reply by Searle).

[15] J.R. Searle, ‘The critique of cognitive reason’, in Readings in Philosophy and Cognitive Science*, eds. A. Goldman*, 833–847, MIT Press, Cambridge, (1993).

The Possibility of Indeterminate Cases of Consciousness and the Ethics of AI

Dr. David Mathers

As we construct more and more sophisticated artificial agents, it becomes possible that we will one day construct an AI which is conscious. This raises a new set of ethical issues concerning not what impact the use of this AI has on humans but rather what impact our use of the AI has on the AI itself. For things which are conscious can objects of moral concerns, since conscious mental states are bearers of moral value. For example, it is bad when human beings, or animals consciously experience pain, and, all things being equal. we have reasons to prevent this from happening. So if we're concerned with ethical issues, it will one day, perhaps in the not-too distant future, be important for us to know whether the artificial agents we are constructing are conscious.

In this paper, I will focus on two more specific questions around artificial consciousness. Firstly, whether there is always a fact of the matter about whether or not a particular agent counts as conscious. And secondly, what duties, if any, we have towards agents where there is no objective fact of the matter as to whether or not they are conscious. In particular, I want to explore the implications of the following possibility: there are real phenomena in the brain which fit distinct and incompatible philosophical or scientific accounts of what consciousness is; further there's no fact of the matter which of these phenomena counts as being consciousness, since each of the phenomena fits our pre-theoretic conception of consciousness equally well. If this is correct, then there will equally be no fact of the matter about what, exactly, you have to do in order to build an AI which is conscious. Rather, when an AI is such that it's mental states count as conscious according to some, but not all of the tied, equally good theories of 'consciousness', there will simply be no fact of the matter about whether it is conscious or not, and hence no fact of the matter, for it's individual mental states, about whether or not they are conscious. Here, I will first motivate the claim that facts about consciousness can go indeterminate in this way. I'll then discuss whether, given such indeterminacy, there is equally no fact of the matter over whether such an artificial agent would be an appropriate object of moral concern (given that being conscious is plausibly a necessary condition on the latter), and no fact of the matter as to whether it's mental states are potential bearers of moral (dis)value.

In the first part of the talk, I will motivate the no fact of the matter claim itself, by motivating a claim I call '**Theory Equality in the Metaphysics of Consciousness' (TEMC)**:

TEMC: Amongst the standard accounts of the metaphysics of 'consciousness' there are multiple, competing and inconsistent, accounts of what consciousness really is; none of these accounts are better than any of the others, because they all do equally well at preserving the (wide) conceptual role of CONSCIOUSNES; it's therefore semantically indeterminate whether mental states which count as conscious according to some but not all of these theories count as cases of consciousness.

I'll motivate TEMC by giving some reasons to think that higher-order theories of consciousness (Carruthers 2016) do better than first-order representationalist theories (Tye 1995) at capturing the theoretical claims to which ordinary thought about consciousness is committed, but worse at capturing the pattern of applications of the term in actual cases which ordinary speakers display. And I'll argue that, given this, it's at least not obvious that higher-order theories do *either better or worse* than first-order representationalist theories, at giving us a 'real definition' of consciousness.

I'll then argue that if this is so, it's plausible that semantic indeterminacy will result, at least if we favour a semantic rather than an epistemic theory of vagueness[1] more generally, and so reject the view that words and concepts always have perfectly precise extensions. In particular, I will sketch an imaginary case where a community's use of a word involves a tension between the actual dispositions subjects have to apply it in particular cases, and their folk theorizing about the necessary conditions on falling under the term in question, which I'll argue is relevantly similar to the tension in our concept CONSCIOUSNESS[2] described in the previous paragraph. I'll then argue that vagueness is a plausible diagnosis in the imaginary case (for both the word meaning, and the concept which figures in the thoughts the word expresses), and that therefore, by analogy, vagueness is also a plausible diagnosis in the case of CONSCIOUSNESS. On the assumption that vagueness is semantic, there will therefore be semantic indeterminacy involving CONSCIOUSNESS (and 'consciousness' and its equivalents in other natural languages). In particular when a mental state counts as conscious by one but not both of the best first-order, and the best higher-order representationalist theories, it will be vague whether that mental state is conscious. And if we design an AI which has at least some representational states that fit one of the definitions, but none that fit both, it will simply be vague whether we have succeeded in designing a conscious AI.

I'll then further defend the idea that nothing fixes an extension of CONSCIOUSNESS more precise than one that is vague in the above manner, by looking at the most plausible candidate ways in which a more precise extension might come to be fixed, and arguing that in each case, it's non-obvious that the extension can get precisified in this way. Firstly, I'll look at the suggestion that out of the relevant higher-order and first-order functional properties that the relevant theories identify the property *being conscious* with, only one of these properties is causally responsible for triggering applications of CONSCIOUSNESS. I'll argue that any straightforward causal account of how the extension of CONSCIOUSNESS gets determined looks like it will simply beg the question against higher-order theories, and is therefore dubious given the plausibility of the latter, and should probably be rejected. This does not prove that there's no more subtle causal account available which would help narrow down the extension of CONSCIOUSNESS, but the failure of any simple causal account puts the burden of proof on anyone who claims an adequate complex account exists. I'll then consider whether choosing one or the other of the higher-order and the first-order properties might lead to CONSCIOUNSESS counting as a more natural kind, than choosing the other. On some views of meaning, natural kinds are 'reference magnets'[3], and so if one of the two properties picked out by the opposing theories was more of a natural kind than the other, this might pull the extension of CONSCIOUSNESS in the direction of that property rather than the other. Here I'll argue that since the relevant properties are both the sorts of properties that might be cited in explanations in cognitive

[1] On the relevant technical notion of vagueness see Sorensen (2012) and Williamson (1994).

[2] I follow the Fodorian convention of using capitalization to get words to denote concept themselves, rather than the properties the concepts pick out.

[3] See Lewis (1983) and Sider (2011, ch.3.2) on natural kinds and reference magnetism.

science, there is no positive reason to think either would be more of a natural kind than the other. Finally, I'll consider the proposal that being conscious is either the disjunctive property of counting as conscious by one or other of the theories, or the conjunctive property of counting as conscious by the standards of both theories. I'll argue that these alternatives are to be rejected on naturalness grounds, and that, in any case, the first does not really accommodate the *a priori* theoretical case for a HOT-theory better than does the standard first-order alternative, whilst the second does no better than standard HOT-theories in agreeing with the actual applications of CONSCIOUSNESS that we are prepared to make.

I'll also sketch some further ways in which TEMC might come out true, even if this particular argument fails, and briefly state why I think it's far from obvious that none of these possibilities are realized. In particular, theories of consciousness are generally proposed precisely *because* they seem to both get the obvious cases correct, and to fit with folk theory of what consciousness is. I'll argue that this itself, plus the lack of convergence in the literature on any particular theory, suggests that there may be no *single* theory of consciousness which best performs the task of both classing the obvious, paradigm cases of conscious mental states as conscious, and fitting with folk theory about consciousness. If the above case for semantic indeterminacy given a tie between higher-order and first-order theories of consciousness is correct, vagueness will follow from this lack of a clear winner.

Having motivated TEMC, I'll then go on to discuss the ethical implications if TEMC is true, in particular whether it forces us to say that when agents, artificial or otherwise, count as conscious by some but not all reasonable definitions of 'conscious', there is no fact of the matter about whether those agents are appropriate subjects of moral concern.

I'll first argue that whether or not there is an objective fact of the matter about whether such *agents* are appropriate subjects of moral concern, turns on the following. Whether there's a fact of the mater about whether *individual mental states* which count as conscious by some but not all reasonable criteria can be bearers of moral (dis) value.

I'll then go on to tackle the latter question about individual mental states. I'll first argue that if there is no way to treat the borderline cases of consciousness created by TEMC as cases where an enabling condition on possessing value is *partially* met, then we will have to say that there's no fact of the matter about whether or not mental states which count as conscious by some but not all of the reasonable theories of consciousness, are bearers of value. The argument here is that if there's no way to see those mental states as meeting the condition for being bearers of (dis)value less than clearly conscious mental states, but more than clearly unconscious ones, then there's no reason to assign them an ability to be bears of (dis)value to a degree intermediate between that of clearly conscious and clearly unconscious mental states, but that it would be arbitrary to assign them the same value-bearing ability as either the former or the latter. Hence, the only option left is to say that it's indeterminate whether they are bearers of value. In the light of this, I'll then discuss the prospects for making sense of the idea that mental states in the indeterminate zone partially but not fully meet the enabling criteria for being a bearer of value that consists in being conscious.

In particular, I'll briefly explore a model for such a view. On this model, each such mental state gets assigned a real number greater than 0 and less than 1, representing the degree to which it meets the condition of being 'conscious', and a number representing the (dis)value it would have, were it a clear case of a conscious mental state, and we then treat it's (dis)value as the result of multiplying the second number by the first. And the real number representing how close to consciousness each mental state is is given by taking the number of the 'tied' theories which count the mental state as conscious over the total number of 'tied' theories. I'll discuss some problems for this model:

Firstly, I'll discuss a problem that arises if TEMC is true, but some of the tied theories seem very slightly variants of each other, whilst other seem to differ more drastically. Consider, for example, a view on which the following three theories are 'tied'; a higher-order theory on which the property *being conscious* is identical to being a representational state that's the subject of a higher-order thought, a first-order representational theory on which *being conscious* is a matter of being a mental state who's representational content is available to many different subsystems within a cognitive architecture[4]; and a first-order representational theory, identical to the last, except that it also demands that the relevant representational content is nonconceptual. Arguably, given how similar the second and third theories are, agreement with both of them ought to count less in terms of closeness to being determinately conscious, than agreement with one of them and the far-more dissimilar higher-order theory. But the way of dealing with indeterminate cases of CONSCIOUSNESS proposed in the previous paragraph, is incompatible with this. Further, it's not clear how to amend it to take account of the reduced importance of counting as conscious by the lights of two highly similar theories, relative to two more dissimilar ones.

Secondly, the method suggested for dealing with indeterminate cases involves treating values as numbers that can be multiplied. But on some reasonable moral views, this cannot be correct. In particular, on some moral views, it can be neither true that two things are exactly equally morally valuable, nor true that one is more valuable than the other, because the values involved are not fully comparable. Given this, values cannot be represented by real numbers, since for any pair of real numbers, *n* and *k*, either *n*=*k*, *n*>*k*, or *k*>*n*. So the method proposed will not work on some reasonable moral views.

I won't reach any firm conclusions in such a short talk about whether these problem can be overcome, but I will suggest some connections to the literature on the relationship between vagueness, and the notion of being true to a degree[5].

References

Baars, Bernard 1997: 'In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness', *Journal of Consciousness Studies*, Vol.4, No.4, pp.292-309.

Carruthers, Peter 2016: 'Higher-Order Theories of Consciousness', *Stanford Encyclopedia of Philosophy*, available at https://plato.stanford.edu/entries/consciousness-higher/

Lewis, David K. 1983: 'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, Vol.61, No.4 pp.343-77.

Sider, Theodore 2011: *Writing the Book of the World*, Oxford University Press, Oxford.

[4] See Baars (1997).

[5] On Vagueness and degrees of truth see (sc.4) of Sorenseen (2012) and (ch.s4+5.5) of Williamson (1994).

Sorensen, Roy 2012: ‘Vagueness’ in *Stanford Encyclopedia of Philosophy*, available at https://plato.stanford.edu/entries/vagueness/

Tye, Michael 1995: *Ten Problems of Consciousness*, MIT Press, Cambridge Mass.

Williamson, Timothy 1994: *Vagueness*, Routledge, London.

Can the *g* Factor Play a Role in Artificial General Intelligence Research?

Davide Serpico[1] **and Marcello Frixione**[2]

Abstract. In recent years, a trend in AI research has started to pursue human-level, general artificial intelligence (AGI). Although the AGI framework is characterized by different viewpoints on what intelligence is and on how to implement it in artificial systems, it conceptualizes intelligence as flexible, general-purposed, and capable of self-adapting to different contexts and tasks. Two questions remain open: a) should AGI projects simulate the biological, neural, and cognitive mechanisms realising the human intelligent behaviour? and b) what is the relationship, if any, between the concept of general intelligence adopted by AGI and that adopted by psychometricians, i.e., the *g* factor? In this paper, we address these questions and invite researchers in AI to open a discussion on the theoretical conceptions and practical purposes of the AGI approach.

1 INTRODUCTION: THE AGI HYPOTHESIS

The dream of the first generation of AI researchers was to build a computer system capable of displaying a human-like intelligent behaviour in a wide range of domains. Since human intelligence is highly flexible with respect to different tasks, goals, and contexts, making the dream come true would have required developing a general-purpose thinking machine.

In spite of some initial success (e.g., Newell and Simon's General Problem Solver [1]), the attempts of researchers did not result in a domain-general AI. What they achieved was, rather, the development of highly specialised artificial systems that behave intelligently in narrow domains, namely "narrow AI". These kinds of artificial systems can carry out domain-specific intelligent behaviours in specific contexts and are, thus, unable to self-adapt to changes in the context as general-intelligent systems can do [2-4].

The realisation of a human-level artificial intelligence has seemed unfeasible to many scholars until recent years. However, in the last two decades, the AI community has started to pursue the goal of a human-level artificial general intelligence (AGI). This is attested by several conferences, publications, and projects on human-level intelligence and related topics [4-5]. Although these projects point to many different directions to be followed by AGI research, they represent a new movement towards the concrete realisation of the original dream of a "strong AI".

Two important movements intertwined with AGI emphasize the importance of the simulation of the human mind. The first, known as Biologically Inspired Cognitive Architectures (BICA), aims to integrate many research efforts involved in creating a computational equivalent of the human mind. The second, which has been initially proposed during the First Annual Conference on Advances in Cognitive Systems (Palo Alto, 2012), aims to achieve the goals of the original AI and cognitive science, that is, explaining the mind in computational terms and reproducing the entire range of human cognitive abilities in computational artefacts [3].

As we mentioned, the AGI community understands general intelligence as the ability, displayed by humans, to solve a variety of cognitive problems in different contexts. Thus, nearly all AGI researchers converge on treating intelligence as a whole: indeed, intelligence appears as a total system of which one cannot conceive one part without bringing in all of it [5-6]. Goertzel [4] delineates the core AGI hypothesis as following: the creation and study of a synthetic intelligence with sufficiently broad scope and strong generalization capability is qualitatively different from the creation and study of a synthetic intelligence with significantly narrower scopes and weaker generalization capability.

What is general intelligence? How can it be implemented in artificial systems? In order to address these questions, it is necessary to open a discussion on both theoretical and practical issues in AGI research.

In this paper, we aim to clarify what relationship exists, if any, between the concept of human general intelligence and the AGI hypothesis. General intelligence was first conceptualised in the early twentieth century within psychometric research. Remarkably, as we shall show, psychometrics is quite a different kind of psychological science than the one traditionally tied to AI, that is, cognitive science. We shall argue that AGI researchers cannot safely rely on the psychometric concept of general intelligence and should, rather, look at intelligence as emerging from several distinct biological and cognitive processes.

In Section 2, we analyse different positions about whether AGI research should emulate or simulate human intelligence. Since many AGI projects are inspired by psychological, neuroscientific, and biological data about human intelligence, scholars in AI should care about the psychometric theory of general intelligence, its promises and perils. In Sections 3 and 4, we summarise the fundamental aspects of such a theory by emphasising the widespread disagreement about the existence of general intelligence. In Section 5, we outline important implications for contemporary research on Artificial General Intelligence.

[1] School of Philosophy, Religion and History of Science, Univ. of Leeds, Leeds, LS2 9JT, UK. Email: D.Serpico@leeds.ac.uk

[2] Dept. of Antiquity, Philosophy, and History, Univ. of Genoa, Via Balbi, 2, 16126, Genoa. Email: marcello.frixione@unige.it

2 EMULATING OR SIMULATING GENERAL INTELLIGENCE?

Human intelligence is defined by psychometricians as a domain-general cognitive ability, namely, the *g* factor (see Section 3 for details). Wang and Goertzel [5] have rapidly dismissed any connection between AGI research and the psychometric concept of general intelligence. According to them, projects in AI are not interested in the psychological description of human intelligence, if not in a weak sense.

However, this conclusion seems to be, at best, premature. Indeed, some attempts in AGI research have encompassed a notion of intelligence that should be evaluated through the lenses of empirical findings. Since general intelligence represents to many psychologists, neuroscientists, and geneticists the most important and well-studied aspect of human psychology [7, 8], we cannot see any strong argument against the possible role of the *g* factor in (at least some) research in AGI. Let us see why.

An AGI project can aim to either emulate or simulate human intelligence. In the case of emulation, an artificial system will display a human-like intelligent behaviour regardless of details about its realisation or implementation.[3] In the case of simulation, instead, an artificial system will display general intelligence not only on a behavioural level but also on a mechanistic and processing level. In other words, human-level artificial intelligence is realized by underlying mechanisms which are analogue to those realising human intelligence. The former case likely represents the notion of AGI that Wang and Goertzel [5] have in mind. Our targets are, instead, examples of AGI research characterised by the latter approach.

Before analysing this approach in more details, it is worth considering that whether AGI projects should emulate or simulate human intelligence, and the relationship between BICA and AGI, are controversial topics. Franklin and colleagues [3, 10] agree that an AGI agent may be successfully developed by using an architecture that is not biologically inspired. However, they argue, the goals of AGI and BICA are essentially equivalent. Indeed, AGI hopes to solve the problem already solved by biological cognition, namely, to generate adaptive behaviour on the basis of sensory input. Since biological minds represent the sole examples of the sort of robust, flexible, systems-level control architectures needed to achieve human-level intelligence, copying after these biological examples—as BICA projects do—represents a valuable strategy.[4]

Wang [11] disagrees with this point of view. According to him, in a broad sense, all AI projects take the human mind as the source of inspiration. Nonetheless, few AI researchers have proposed to duplicate a human cognitive feature without providing a reason why this is needed—consider that computers and human beings are different from each other in many fundamental aspects. Therefore, the important decision for an AGI project is *where* to be similar to the human mind and *why* this similarity is desired.

Our aim is not to take a side in this controversy, but rather to show that some AGI projects are, in fact, inspired by empirical data on human intelligence. Hassabis and colleagues' review [12] provides several examples of how neuroscience has inspired both algorithms and artificial architectures. Moreover, neuroscience seems to be able to provide validation of already existing AI techniques as well: if a known algorithm is found to be implemented in the brain, then that is strong support for its plausibility as an integral component of an intelligent system. In this view, brain studies have helped developing AI architectures enlightening the functioning of central aspects of intelligence such as learning, attention, and memory.

A dialogue between neuroscientific and AI research seems to be largely welcomed within the AGI community. A survey conducted by Muller and Bostrom [13] highlights how, according to many researchers, a human-level AI will likely be achieved by means of research approaches tying AI to neuroscience—e.g., Integrated Cognitive Architectures, Computational Neuroscience, and Whole Brain Emulation. Of course, the commitment to the simulation of psychological, cognitive, and biological aspects of human intelligence is exerted in many ways. Let us see some examples.

Some projects belonging to BICA and AGI's agendas (e.g., SyNAPSE, HTM, SAL, ACT-R, ICARUS, LIDA, the ANNs, the Human Brain Project, and the Large-Scale Brain Simulator) are interested in various aspects of the human general intelligence and accept, though to different degree, that simulating the human brain's structure can be promising for AI research [3, 14-20].

Further, various researchers are inspired by the ontogenetic and phylogenetic aspects of human intelligence and suggest that we should simulate the same facilities for learning that human infants have or the evolutionary trajectory of intelligence [9, 21, 22].

Lastly, Wang [6] suggests that an AGI system may require a single mechanism capable of reproducing the general-purpose, flexibility, and integration of human intelligence. Accordingly, a general intelligent system should comprise both domain-specific and domain-general sub-systems: while the existing domain-specific AI techniques are considered tools for solving specific problems, the integrating component is general, flexible, and can run the various domain-specific programs. In the proposed architecture, i.e., the NARS, reasoning, learning, and categorisation represent different aspects of the same processes. This approach, highlighting the relationship between general intelligence and a hypothetic domain-general mechanism, is particularly interesting for our purposes. Indeed, as we shall explain shortly, the psychometric theory of human intelligence draws on similar intuitions.

Is human intelligence related to a single, general cognitive mechanism? How can general intelligence emerge from the complexity of the human brain? To the extent that AGI researchers aim to reproduce human intelligence on a mechanistic and processing level, they should care about these questions, which are typically addressed by empirical research on human intelligence. In the next two sections, we review the psychometric theory of general intelligence and ask whether AGI and BICA projects can safely rely on it.

[3] Here, behavioural assessments, such as Turing's test and Nilsson's employment test, can address whether we have achieved a human-level artificial intelligence: in brief, systems with true human-level intelligence should be able to perform human-like tasks [9].

[4] For instance, since mind and brain are strictly related, the LIDA's theoretical model proposed by the authors seeks to reproduce it *in silico.*

3 THE THEORY OF HUMAN GENERAL INTELLIGENCE

The concept of general intelligence was born in the early twentieth century, in parallel with the rise of the psychometric tradition. The central aim of psychometricians is to develop methodologies capable of assessing and quantifying intellectual differences among peoples, i.e., IQ tests. Over the last century, these tests served a variety of purposes, ranging from educational to clinical ones, but they played a central role in empirical research as well (e.g., in behavioural genetics and neuroscience).

The concept of intelligence is generally related to a wide range of psychological aspects and adaptive capabilities (e.g., learning, knowledge, social skills, and creativity; see [23]). By contrast, psychometricians have mainly focused on the cognitive abilities mostly involved in the solutions of the problems included in IQ tests (e.g., mathematical, linguistic, logic, and visual-spatial abilities). Thus, general intelligence represents a theoretical construct related to these cognitive domains.

In order to clarify the nature of general intelligence, psychometricians generally refer to what Charles Spearman [24] called the general factor of intelligence or *g* factor. Remarkably, there are two different ways of understanding *g*: on the one hand, the psychometric *g*; on the other hand, the neurocognitive *g*. Let us see them one by one.

From a psychometric point of view, the *g* factor is related to the so-called positive manifold: individuals who show good performance on a given task will tend to show good performance also in other tasks. In other words, intelligence measurements are positively intercorrelated both in different cognitive domains and different individuals. Spearman understood that scores of a battery of tests tend to load on one major factor regardless of their domain. He employed factor analysis to identify this factor. The *g* factor, as Spearman called it, is a latent variable which summarises the typical correlation matrix of intelligence test scores.

What is the meaning of the psychometric *g*? Factor analysis can be understood as a procedure of "distillation" capable of identifying a factor reflecting the *variance* that different intellectual measures have in common. In this sense, the *g* factor explains ~40 percent of tests' variance. Thus, it reflects *individual differences* in performance in intellectual tasks [7, 25]. This interpretation of *g* can hardly find room in neuroscientific research: indeed, the psychometric *g* does not represent a concrete neurocognitive mechanism, but rather an abstract entity or a property of a population of individuals (see [26] for similar concerns].

From a neurocognitive point of view, the story is different. In neuroscience, the *g* factor is understood as a domain-general cognitive ability that characterises human beings [27]. In this respect, it represents the fundamental mechanism underlying general intelligence. However, the meaning of the neurocognitive *g* is still unclear. When Spearman tried to clarify the nature of intelligence, he described *g* as a form of mental energy. Successive researchers have tried to reduce *g* to some neurocognitive properties of the human brain, e.g., working memory, processing speed, and neural efficiency (see Section 4 for further details).

The reliability of the psychometric *g* is generally accepted as the positive manifold represents a stable empirical phenomenon. By contrast, several concerns have been raised on the neurocognitive interpretation of *g*.

In recent decades, many neuroscientists have come to represent the central detractors of the concept of general intelligence. Most of them have developed non-general conceptions of human intelligence; others have interpreted *g* as a mere statistical artefact. In the next section, we analyse the controversial role of the general factor of intelligence in neuroscientific research.

4 A NEUROSCIENTIFIC VIEW ON THE G FACTOR

Is there any evidence of the existence of *g* in the human brain? Is human intelligence general or not? Since psychometrics and cognitive science met a few decades ago, these questions divide scholars for both empirical and theoretical reasons.

From an empirical point of view, the pro-*g* scholars have tried to reduce *g* to neurocognitive constructs, often assumed as reliable and hence suitable to make sense of *g* in neuroscientific terms. Associations have been found, for instance, between IQ and processing speed, working memory, problem-solving, meta-cognition, attention, associative learning, glucose metabolic rates, electrocortical activity, and brain size [28-30]. However, to find reliable associations between *g* and these variables has not been easy at all: replicability rates are often low and spurious correlations ubiquitous. Moreover, the associations between *g* and other aspects of the human brain are often considered to be theoretically inconsistent or, at best, weak [31, 32].

From a theoretical point of view, the pro-*g* scholars have developed theories of intelligence aimed at reconciling neuroscientific and psychometric approaches. For instance, the Minimal Cognitive Architecture Theory [33] aims to match the psychometric view with developmental theories of intelligence and with the modular theory of mind. The Parieto-Frontal Integration Theory [30, 34], in turn, aims to locate the *g* factor into the human brain, i.e., in the parietal and frontal regions.

Although these models represent interesting attempts, many scholars believe there is no room for general intelligence in contemporary neuroscience. Indeed, most contemporary theories of intelligence do not include the *g* factor within the human cognitive architecture and do not identify a single general mechanism capable of summarising individual performances as a global test-score such as IQ. Rather, several aspects of biology and cognition are invoked. Renowned examples are the theory of Multiple Intelligences [35], the PASS model [36], and the Multiple Cognitive Mechanisms approach [37]. All these theories appeal to the role of several distinct cognitive processes to explain the human intelligent behaviour.

If there is no a general mechanism such as *g* in the human brain, why then the positive manifold? Some scholars have recently provided valuable explanations of the empirical correlations among IQ-tests performance without invoking a general underlying mechanism. According to these proposals, the psychometric *g* is supported by multiple, interacting mechanisms that become associated with each other throughout the course of development. For instance, the mutualist model, proposed by Van der Maas and colleagues [37], recognises that the positive manifold is a robust empirical phenomenon, but advances an explanation based on a developmental model involving the relationships between cognitive processes. The mutual influence between these processes gives rise to the positive manifold but rules out *g* as a single, latent variable. According to the architects of this model, there is nothing wrong with using the *g* factor as a summary index as long as we do not assume that this variable relates to a single underlying process.[5]

To summarise, cognitive neuroscientists often deny the existence of the neurocognitive *g* and, thus, suggest that general intelligence does not represent a valuable posit for understanding human cognition (see also [38]). The disagreement about the existence of the *g* factor can be clarified by considering the theoretical gap between psychometrics and cognitive science. Since the birth of cognitive psychology, cognitive scientists have focused on the functional-structural segmentation of the human mind. Thus, in a neurocognitive perspective, mental abilities and cognitive processes cannot be considered properties of the brain taken as a whole: rather, they are implemented by specific brain-areas and populations of neurons (for instance, the modularity of mind hypothesis relies on this assumption). This conclusion is sometimes agreed by researchers in AGI as well. For instance, Goertzel [4] has contrasted the conception of general intelligence with approaches looking at the various competencies humans display (see the list of competences assembled at the 2009 AGI Roadmap Workshop [2]).

In the last section, we explore some implications for research in AGI.

5 CONCLUSIONS: IMPLICATIONS FOR ARTIFICIAL INTELLIGENCE RESEARCH

The quest for the nature of human intelligence, involving its generality and its architecture, remains open. However, it stands to reason that general intelligence cannot be safely understood as a real biological entity. Rather, we can describe it as a behavioural, emergent phenomenon due to the causal interaction between many aspects of the neurocognitive development. Accordingly, intelligence seems to be a term imported by everyday life that clusters together distinct cognitive processes, autonomous to a certain extent both in developmental and evolutionary terms.

What does this imply for AI researchers who adopt a generalist view of intelligence? Two things, at least. First, the generalist conception of intelligence, if adopted in AGI and BICA, risks inheriting the weaknesses of its relative in the human domain, the psychometric one. Artificial systems inspired by such a theory can well turn out to be less human-oriented than other, classical ones, such as the so-called narrow AI systems. Second, implementing some sort of general-purpose mechanism in artificial systems to emulate the human intelligent behaviour—as Wang [6], among others, suggests—may not be the right strategy.

It is worth noting that, in general, a psychometric-like view of intelligence does not play a central role in AI. Indeed, most contemporary artificial architectures do not assume that a human-level intelligence necessarily requires a single generative mechanism. Rather, intelligence is understood as emerging from many underlying aspects—an interpretation with which, as we have shown, many cognitive neuroscientists agree. At the same time, almost any scholar would recognise that the classical narrow approach to AI is unsuccessful in shifting towards a human-level intelligence.

So, what there is between specialised artificial systems and a single domain-general mechanism? Is there any intermediate level to work on? Essentially, these are the questions AGI researchers aim to address (see [14, 39]). In other words, AGI researchers are asked to develop lower-level, specific-purposed systems capable of generating higher-level networks of processes and interactions. These networks would arguably realise general intelligence on the behavioural level. Indeed, intelligence represents a systemic and dynamical property of complex systems.

Unfortunately, even complex cognitive architectures, such as SOAR and ACT-R, are characterised by both technical and epistemological problems (see e.g., [14, 40, 41]). Neuroscientific theories of intelligence can help AI by providing a meaningful explanation of human neurocognitive development. Nevertheless, taking up the challenge ultimately depends on the ability of AI researchers to pick up the relevant conceptions of what an intelligent system is. In this sense, a discussion on general intelligence in AI seems to us, at present, inevitable.

Can the *g* factor play a role in AGI research, after all? In light of our discussion, the answer can be either positive or negative. Roughly, the answer depends on what the aims of AGI are. A weak or instrumental notion of *g*, like the psychometric *g*, can play a role in AGI projects characterised by an emulative approach, where the goal is reproducing a human-level intelligence regardless of details about its neurocognitive or biological architecture. Here, the psychometric *g*, as assessed by IQ tests, might help to evaluate the intelligent behaviour of artificial systems besides other behavioural tests—e.g., Turing and Nilsson's tests.

By contrast, a strong, neurocognitive notion of *g* is involved in the discussion about the composition of human intelligent systems, the causal interactions among parts, and how to artificially reproduce these aspects. In this respect, the possible role of the *g* factor in AGI research depends on empirical data in neuroscience. As we have argued in this paper, this role of *g* is dubious.

As we noticed, AGI research encompasses different viewpoints on what intelligence is and on what the purposes of a human-level AI are. While many authors in AGI are cautious about their assumptions, others believe it is not enough to merely emulate the intelligent behaviour. Rather, in this view, artificial systems should simulate the mechanisms and processes that make humans intelligent the way they are. For these approaches, where theories and data adopted by cognitive neuroscientists play an important role, we invite cautious about the commitment to the concept of general intelligence. As Goertzel [4] notices, brain sciences are advancing rapidly, but our knowledge about the brain is extremely incomplete. Seemingly, relying on a controversial theory of human intelligence, such as the psychometric one, can be perilous for AI research.

REFERENCES

[1] Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: symbols and search. The Tenth Turing Lecture. *Communications of the Association for Computing Machinery, 19*.

[2] Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., ... & Shapiro, S. C. (2012). Mapping the landscape of human-level artificial general intelligence. *AI magazine*, *33*(1), 25-42.

[5] See [2] for developmental approaches in AGI.

[3] Franklin, S., Strain, S., McCall, R., & Baars, B. (2013). Conceptual commitments of the LIDA model of cognition. *Journal of Artificial General Intelligence*, *4*(2), 1-22.

[4] Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, *5*(1), 1-48.

[5] Wang, P., & Goertzel, B. (2007). Introduction: Aspects of artificial general intelligence. In *Proceedings of the 2007 conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* (pp. 1-16). IOS Press.

[6] Wang, P. (2004). Toward a unified artificial intelligence. In *AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Research and Systems* (pp. 83-90).

[7] Jensen, A. R. (2002). Psychometric g: Definition and Substantiation. In R. Sternberg & E. Grigorenko (Eds.), *The General Factor of Intelligence. How General is it?* (pp. 39-54). Mahwah: Lawrence Erlbaum.

[8] Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. N. (2013). *Behavioral Genetics* (6th ed.). New York: Worth Publishers.

[9] Nilsson, N. J. (2005). Human-level artificial intelligence? Be serious!. *AI magazine*, *26*(4), 68-75.

[10] Strein, S., & Franklin, S. (2013). Authors' Response to Commentaries. In H., Dindo, J., Marshall, & G., Pezzulo (2013). Editorial: Conceptual Commitments of AGI Systems. *Journal of Artificial General Intelligence*, *4*(2), 48-58.

[11] Wang, P. (2013). Conceptual Commitments of AGI Projects. In H., Dindo, J., Marshall, & G., Pezzulo (2013). Editorial: Conceptual Commitments of AGI Systems. *Journal of Artificial General Intelligence*, *4*(2), 39-42.

[12] Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245-258.

[13] Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 555-572). Berlin: Springer.

[14] Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review, 111*(4), 1036-1060.

[15] Sun, R., & Naveh, I. (2004). Simulating organizational decision-making using a cognitively realistic agent model. *Journal of Artificial Societies and Social Simulation, 7*(3).

[16] Gunzelmann, G., Gluck, K. A., Van Dongen, H. P. A., O'Connor, R. M., & Dinges, D. F. (2005). A Neurobehaviorally Inspired ACT-R Model of Sleep Deprivation: Decreased Performance in Psychomotor Vigilance. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Meeting of the Cognitive Science Society* (pp. 857-862). Mahwah: Lawrence Erlbaum.

[17] Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience* *7*(2), 153-160.

[18] Chong, H. Q., Tan, A. H., & Ng, G. W. (2007). Integrated cognitive architectures: a survey. *Artificial Intelligence Review*, *28*(2), 103-130.

[19] Izhikevich, E. M., & Edelman, G. M. (2008). Large-Scale Model of Mammalian Thalamocortical Systems. *PNAS, 105*, 3593-3598.

[20] Jilk, D. J., & Lebiere, C. (2008). SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence 20*, 197-218.

[21] Brooks, R. A. (1997). From Earwigs to Humans. *Robotics and Autonomous Systems*, *20* (2-4), 291-304.

[22] Koza, J. R., et al. (2003). *Genetic Programming IV: Routine Human-Competitive Machine Intelligence.* Norwell, MA: Kluwer Academic Publishers.

[23] Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, *157*.

[24] Spearman, C. (1923). *The Nature of Intelligence and the principles of cognition*. London: MacMillan.

[25] Stankov, L. (2002). g: A Diminutive General. In R. Sternberg & E. Grigorenko (Eds.), *The General Factor of Intelligence. How General is it?* (pp. 19-38). Mahwah: Lawrence Erlbaum.

[26] Borsboom, D., & Dolan, C. V. (2006). Why g is not an adaptation: A comment on Kanazawa (2004). *Psychological Review, 113*(2), 433-437.

[27] Burkart, J. M., Schubiger, M. N., & van Schaik, C. P. (2017). The evolution of general intelligence. *Behavioral and Brain Sciences*, *40, 1-24*.

[28] Pretz, J. E., & Sternberg, R. J. (2005). Unifying the Field: Cognition and Intelligence. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and Intelligence* (pp. 306-318). Cambridge, UK: CUP.

[29] Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*(3), 201-211.

[30] Haier, R. J. (2017). *The neuroscience of intelligence*. Cambridge University Press.

[31] Kray, J., & Frensch, P. (2002). A View From Cognitive Psychology: *g*–(G)host in the Correlation Matrix? In R. Sternberg & E. Grigorenko (Eds.), *The General Factor of Intelligence. How General is it?* (pp. 183-220). Mahwah: Lawrence Erlbaum.

[32] Ramus, F. (2017). General intelligence is an emerging property, not an evolutionary puzzle. *Behavioral and Brain Sciences*, *40*, 43-44.

[33] Anderson, M. (2005). Marrying intelligence and cognition. A Developmental View. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and Intelligence* (pp. 268-287). Cambridge, UK: CUP.

[34] Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: converging neuroimaging evidence. *Behavioral and Brain Sciences*, *30*(2), 135-154.

[35] Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic books.

[36] Naglieri, J. A., & Das, J. P. (2002). Practical Implications of General Intelligence and PASS Cognitive Processes. In R. Sternberg & E. Grigorenko (Eds.), *The General Factor of Intelligence. How General is it?* (pp. 55-84). Mahwah: Lawrence Erlbaum.

[37] Van der Maas, H. J. L., Dolan, C. V., Grasman, R. P. P. P., et al. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological review, 113*(4), 842-861.

[38] Serpico, D. (2017). What Kind of Kind is Intelligence?. *Philosophical Psychology, 31*(2), 232-252.

[39] Lieto, A., Bhatt, M., Oltramari, A., & Vernon, D. (2018a). The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research, 48,* 1-3.

[40] Lieto, A., Lebiere, C., & Oltramari, A. (2018b). The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, *48*, 39-55.

[41] Laird, J. E. (2012). *The Soar cognitive architecture*. MIT press.